The Signal and the Noise. Differentiating Stylometric Signals in the Analysis of Literary Texts

Brottrager, Judith

(2020)

The Signal and the Noise. Differentiating Stylometric Signals in the Analysis of Literary Texts

Digital Philology |
Evolving Scholarship in Digital Philology - 01 | 2020

Herausgegeben von
Sabine Bartsch | Evelyn Gius | Marcus Müller | Andrea Rapp | Thomas Weitin

Judith Brottrager

TECHNISCHE UNIVERSITÄT DARMSTADT

DIGITALE PHILOLOGIE DARMSTADT

Evolving Scholarship in Digital Philology

Impressum
Postadresse: Technische Universität Darmstadt
Institut für Sprach- und Literaturwissenschaft
Dolivostraße 15
64293 Darmstadt
Website: www.digital-philology.tu-darmstadt.de
Email: sprachli@linglit.tu-darmstadt.de

Zitierhinweis: Judith Brottrager: The Signal and the Noise. Differentiating Stylometric Signals in the Analysis of Literary Texts. In: Digital Philology | Evolving Scholarship in Digital Philology 01|2020. Darmstadt: TUPrints.

Vorwort der Herausgeber*innen zum Start der Reihe

Mit der vorliegenden Publikation ist der Startschuss gefallen für die neue Open-Access-Schriftenreihe “Digital Philology | Evolving Scholarship in Digital Philology”, die von der Digitalen Philologie am Institut für Sprach- und Literaturwissenschaft der Technischen Universität Darmstadt ins Leben gerufen wird. Sie soll ebenso wie die weitere, im Juli 2020 gestartete neue Reihe – “Digital Philology | Working Papers in Digital Philology” – die Publikationslandschaft der Digital Humanities in Deutschland bereichern und profilieren. Im Vorwort zum ersten Band der Schwesterreihe haben wir bereits eine kurze Einführung zum Selbstverständnis der Heraus-geber*innen und zur inhaltlich-fachlichen Ausrichtung der beiden Reihen gegeben, die hier zur besseren Orientierung nochmals aufgenommen wird.

Im Darmstädter Modell der Digital Humanities betrachten wir Digitalität als integrativen Bestandteil der jeweiligen fachlichen Identität; daher bezeichnen wir unsere Forschungstätigkeiten und auch unsere Studiengänge nicht mit dem übergreifenden Begriff bzw. umbrella term ‘Digital Humanities’, sondern spezifizieren das Feld als ‘Digitale Philologie’ bzw. ‘Linguistic and Literary Computing’. Die Beschäftigung mit Digitalität umfasst dabei sowohl die Aspekte der Materialität und Medialität – also die Befassung mit digitalen Objekten – als auch die Entwicklung und Anwendung digitaler Methoden und Verfahren. Mit seinen vier Professuren und einer weiteren Fachgebietsleitung mit dediziert digital-philologischer Ausrichtung hat das Darmstädter Institut für Sprach- und Literaturwissenschaft ein Alleinstellungsmerkmal – mindestens in Deutschland, aber auch weit darüber hinaus.

Die Herausgeber*innen stellen sich vor:

Sabine Bartsch ist Anglistin und Linguistin mit einem Schwerpunkt im Bereich Korpus- und Computerlinguistik und interessiert sich besonders für Kollokationsforschung, historische Register der Wissenschaftskommunikation, Multimodalität sowie Korpusdesign und -aufbau und Methoden korpusbasierter Analyse.

Evelyn Gius ist digitale Literaturwissenschaftlerin und interessiert sich besonders für Annotation, die narrative Struktur und die Segmentierung von literarischen Texten, die Automatisierung bzw. Automatisierbarkeit von Textanalyse sowie die Wechselwirkungen zwischen computationellen Verfahren und etablierten literaturwissenschaftlichen Methoden.

Marcus Müller ist Linguist und interessiert sich für Korpuslinguistik, digitale Diskursanalyse, Wissenschaftskommunikation, grammatische Variation sowie Sprache in der Kunstkommunikation.

Andrea Rapp ist germanistische Mediävistin und Computerphilologin mit Bibliothekserfahrung und interessiert sich besonders für die Digitalisierung und Erschließung mittelalterlicher Handschriften, die Analyse von Urkunden und Briefen, Varianz und Varietäten, Editionsphilologie, Lexikographie, Annotationsverfahren sowie Forschungs(daten)infrastrukturen.

Thomas Weitin ist digitaler Literaturwissenschaftler und interessiert sich für Modelle, die das Verhältnis des kanonischen Teils der Literaturgeschichte zum great unread sichtbar machen. Sein zweiter Forschungsschwerpunkt liegt in der kognitiven Rezeptionsanalyse.

Während die Working-Papers-Reihe dediziert auf die Publikation von Working Papers, White Papers, Diskussionsimpulsen, Projektberichten und ähnlichen Formaten ausgerichtet ist, bietet die Reihe 'Evolving Scholarship in Digital Philology' hervorragenden Abschlussarbeiten (in der Regel aus einem Master) einen Publikationsort. Die Qualitätssicherung erfolgt auf der Basis bestimmter Bewertungsvoraussetzungen (in der Regel Mindestnote 1,7) und erfordert zudem ein Peer Review von außerhalb des Darmstädter Herausgeberteams bzw. die Auszeichnung durch einschlägige Preise. In beiden Reihen sind deutsch- und englischsprachige Publikationen willkommen.

Band 1 der 'Evolving Scholarship'-Reihe startet mit der vom Fachbereich 02 der TU Darmstadt ausgezeichneten Arbeit von Judith Brottrager: The Signal and the Noise. Differentiating Stylometric Signals in the Analysis of Literary Texts. TU Darmstadt 2019.

Sie wurde im Arbeitsgebiet der Digitalen Literaturwissenschaft verfasst und von Thomas Weitin und Sabine Bartsch betreut. Die Arbeit wurde mit der 2020 neu etablierten Auszeichnung der besten Abschlussarbeiten in den Forschungsschwerpunkten des Fachbereichs ausgezeichnet. Die im Folgenden abgedruckte, von Sabine Bartsch gehaltene Laudatio würdigt die Preisträgerin und ihre Arbeit und gibt einen Einblick in den Inhalt der Thesis.

> Judith Brottrager ist mir seit dem Jahr 2016 bekannt, als sie, damals noch im Bachelorat an der Universität Wien, Kontakt zu uns aufnahm, um sich über den Master-of-Arts-Studiengang Linguistic and Literary Computing an der TU Darmstadt zu informieren, und schließlich zum Wintersemester 2017-18 zu uns nach Darmstadt kam, um diesen Master mit großem Engagement und

Erfolg zu studieren, den sie Ende 2019 mit ihrer in englischer Sprache verfassten Master-Thesis unter dem Titel “The Signal and the Noise. Differentiating Stylometric Signals in the Analysis of Literary Texts” absolviert hat.

In ihrer Thesis beschäftigt sich Judith Brottrager anhand des Forschungsfeldes der Stilometrie mit der Frage, inwiefern und unter welchen Voraussetzungen statistische Untersuchungen in den Computational Literary Studies wissenschaftlich valide Ergebnisse liefern und welche Voraussetzungen hierfür besonders mit Blick auf die Repräsentativität und literaturwissenschaftliche Relevanz und Qualität der untersuchten Textkorpora gewährleistet sein müssen.

In Experimentreihen zu stilometrischen Merkmalen zeigt Brottrager Wege auf, wie stilometrische Untersuchungen literarischer Texte Fehlschlüsse aus empirischen Untersuchungen vermeiden und relevantere Ergebnisse aus großen literarischen Korpora ermitteln können. Sie wendet dabei nicht nur aktuelle Verfahren des maschinellen Lernens und der Netzwerkanalyse an, sondern plausibilisiert ihre Ergebnisse auch auf höchstem literaturwissenschaftlichem Niveau. Ein Gutachten bescheinigt, dass ihr methodisches Vorgehen das Zeug zu einer best practice in den Computational Literary Studies hat.

Mit ihrer Thesis zeigt Judith Brottrager im besten Sinne der Digital Humanities die Stärken einer engen Verzahnung philologischer mit technologischer Kompetenz auf.

Seit Januar 2020 ist Judith Brottrager wissenschaftliche Mitarbeiterin am Fachgebiet Germanistik – Digitale Literaturwissenschaft und arbeitet als Doktorandin im DFG-Projekt “Relating the Unread. Network Models in Literary History” (DFG Priority Programme 2207 “Computational Literary Studies”).

Liebe Judith, es ist mir eine Ehre und eine Freude, Dir heute den Preis für eine herausragende Master-Arbeit im Forschungsschwerpunkt Digital Humanities des Fachbereichs 02 Gesellschafts- und Geschichtswissenschaften überreichen zu dürfen.

Die Herausgeber*innen freuen sich ganz besonders, dass diese preiswürdige Arbeit die neue Open-Access-Reihe eröffnet, die die Leistungen junger Forschender würdigen und sichtbar machen soll. Und schließlich möchten wir uns nochmals bei allen bedanken, die das Vorhaben unterstützt haben und weiterhin unterstützen: natürlich bei unseren

Teams, aber ganz besonders bei den Studierenden der verschiedenen Studiengänge, die sich seit dem Wintersemester 2006-07 auf das Abenteuer der Digital Philology an der TU Darmstadt einlassen, die uns mit ihrem Mut und mit ihrer Neugier immer wieder beeindrucken und die sich auch und gerade im Jahr 2020 verantwortungsbewusst, freundlich und zielstrebig auf den Weg in die Digital Humanities machen. Wir danken weiterhin der Universitäts- und Landesbibliothek, die insbesondere mit dem Team Digitales Publizieren Herausgeber*innen und Autor*innen vorbildlich betreut.

The Signal and the Noise

Differentiating Stylometric Signals in the Analysis of Literary Texts
Master thesis by Judith Brottrager
Date of submission: 28.10.2019

1. Review: Prof. Dr. Thomas Weitin
2. Review: Dr. Sabine Bartsch
Darmstadt – D 17

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Contents

List of Figures

List of Tables

1 Introduction

A few months before I started to write this thesis, an essay by Nan Z. Da (2019) caused some considerable uproar in the Digital Humanities (DH) community, more specifically in the field of computational literary studies (CLS). The essay is concerned with problems the author detected in recent studies from the field, ranging from methodological flaws to a lack in scientific rigour. Several members of the community swiftly offered passionate rebuttals to Da's claims (see Eve 2019, Bode 2019, Beausang 2019, Herrmann et al. 2019) and after some intense arguments on Twitter, the discussion subsided quickly. The strong emotionality of this discourse made it impossible to gain something from this situation. This was partially caused by Da's sometimes harsh rhetoric, but also by the overly defensive attitude of some of her respondents. Because although some points raised by Da have been addressed and successfully refuted (see above), some of the critiques she raised are valid and crucial for further developments in CLS. In the following, I will address some of the issues raised by Da and will then go on to outline how her essay will influence my thesis.

Da states, for example, that she tried to reproduce several recent research projects, but often could not find any accompanying scripts or only partial or corrupted versions (2019, 602, footnote 2; 605). This point is not only valid but also immensely vital for a productive approach to CLS. As literary scholars and humanists, we are used to discuss methods and methodological concerns, but still have to learn to make all our material available in a way that enables an easy and transparent reproduction. This means that both code and data have to be made available online. Although this mantra has been a part of the DH and the CLS for several years now, it often seems that even large-scale projects struggle to find a suitable strategy for the publication of this crucial material. Providing thousands of lines of code without extensive commentary hinders a critical debate of results, but also makes it more difficult for beginners in the field to understand more elaborate projects.

In addition to this very practical point, Da (2019, 604) points out that most papers she studied show "[o]versights in implementation; lack of robustness, precision, and recall;

and less than ideal measurements". These problems, which she detects generally in many forms of data-mining (2019, 604), include an unawareness of which tools and methods are best suited for a chosen approach. Similar to more traditional approaches to literary studies, CLS scholars have to be able to argue convincingly why a certain method is suitable for a research question and which limitations are caused by choosing this method. Especially in a field which values methodological critique, it is interesting to see that the choice of method sometimes seems rather motivated by availability and convenience than by suitability.

Da's main point of criticism is, however, that the papers she examined divide into what she calls "no-result papers", i.e. "those that haven't statistically shown us anything" and "papers that do produce results but that are wrong" (2019). The papers are, according to Da (2019, 605),

> more or less all organized the same way, detecting patterns based in word count (or 1-, 2-, n-grams, a 1-gram defined as anything separated by two spaces), to make one of six arguments: (1) the aboutness of something; (2) how much or how little of something there is in a larger body, (3) the amount of influence that something has on something else; (4) the ability of something to be classified; (5) if genre is consistent or mixed; and (6) how something has changed or stayed the same.

These six main research questions are examined and analysed by measuring and statistically representing overlapping vocabulary, by compressing these measures into models and by testing these models (2019, 605). Statistical tests are employed to attempt to indicate causation, even though the "explanation of said causation/correlation through fundamental literary theoretical principles are usually absent as well" (Da 2019, 605).

A lot can be gained by critically engaging with the points raised: First, the issue of feature selection certainly needs more exploration. It is true that most CLS approaches work with counts of n-grams and that this has an enormous impact on the range of research questions that can be addressed in this way. Particularly rare words, for example, are not likely to influence a stylometric analysis, as they are either on a too low rank to be considered in an most frequent word (MFW) approach and/or are eliminated in a so-called culling process.[1] A research project which relies on these words would need to find another strategy for feature selection. Generally, the implications of feature selections are disputed: Even though there are some theories about which size of feature vector is more suitable for one research question than for others—high frequency words are, for instance, often interpreted as markers of authorship (cf. Mosteller and Wallace 1963,

[1] Culling is a word list manipulation; "the culling values specify the degree to which words that do not appear in all the texts of a corpus will be removed" (Eder, Rybicki, Kestemont, and Pielström 2019, 47).

Burrows 2002, Hoover 2001, Craig and Kinney 2012)—a definitive answer has still to be found. The overall tendency for CLS projects to use MFW feature lists seems to be due to a combination of factors: Feature counts are easy to extract and to compare. They hardly need any pre-processing, as no part-of-speech (PoS) tagging or similar processes have to be employed. Some even argue, as I will discuss in the next chapter, that using more elaborate features does not make a considerable difference.

Second, Da's eloquent case against "no-results papers" can be linked to what she later says about hypothesis testing. She defines "no-results papers" as "papers that present a statistical no-result finding as a finding" (2019, 607). These findings are caused by answering a question with the wrong model, or, to put it differently, by selecting an uninformative and ill-fitting null hypothesis. Such a hypothesis, as, for example, "most frequently used words don't change / most frequently used words do change" (Da 2019, 618), can be tested rigorously, but will still lead to incorrect conclusions.

Third, Da's critique of "papers that do produce results but that are wrong", or, as she also puts it, "papers [that] draw conclusions from its findings that are wrong" (2019, 607) can again be linked back to the fact that theories and interpretations are based solely on word counts without taking into account any additional linguistic features. "Word frequencies and the measurement of their differences are", so Da (2019, 611), "asked to do an enormous amount of work, standing in for vastly different things". This should not mean that word frequencies cannot be used in any analysis, but that using them as features comes with implications. These implications heavily influence whether the feature can be used to describe a certain phenomenon—a consideration which once more can be associated with the formulation of a sound null hypothesis.

Forth, Da raises the question whether CLS papers sufficiently connect their findings to literary principles and theories. This claim, coming from a more traditionally working literary scholar, should be understood as a reminder that CLS is more than data science or text mining. Contextualising the research object, i.e. literary texts, in a theoretical framework can help to formulate a suitable hypothesis, but is also vital for the detection of limitations and implications caused by the data.

I have decided to start my thesis with this lengthy discussion of Da's essay because it presents an outsider's view which can serve as a guideline for how to avoid common flaws in CLS projects. Moreover, this critical discussion should set the thesis' tone, as I will try to examine what computational approaches to literary studies can offer and where, to quote Da for a last time, "the threshold of optimal utility" (Da 2019, 639) lies. For this purpose, I have chosen another controversial text as the title and motto of my thesis: *The Signal and the Noise* by Nate Silver (2012). Silver and his statistical analysis of signals, i.e.

sound statistical predictors, and noise, i.e. random observations that tend to obscure the signal, are far from what I am planning to do in this thesis. Silver discusses the prediction of election outcomes, earthquakes, and the weather and describes how valid prognoses are formulated; I will examine literary texts and will try to find out which categories impact the closeness and distance of individual texts in stylometric analysis. Nevertheless, Silver raises a point that seems also very crucial in the computational analysis of literary texts: When interpreting data—be it election polls or distance measures between literary texts—human interpreters display ”almost hyperactive pattern-recognition skills” and ”see patterns where there aren't any” (2012, 277). This pattern detection is ubiquitous in CLS: A distance measure is considered good when its employment leads to a clustering of texts by the same author, even though it is not clear why the distance measure is so productive (cf. Büttner et al. 2017); when a clustering shows predominantly male or female clusters, it is claimed to be due to the influence of a gender signal (cf. Rybicki 2016). The fact that these patterns are easily detectable does, however, not mean that they are caused by the interpretation of noise, but that despite being seemingly easy to decipher, their interpretation needs to be based on a contextualisation of the analytical model.

Considering both the points raised by Da and the implications of Silver's discussion of signal and noise, it is the aim of my thesis to critically engage with stylometric analysis on the micro-level of feature selection and parameter settings, as well as on the macro-level of corpus subsetting. Additionally, I will employ different methods, ranging from descriptive statistics to super- and unsupervised learning, to show their advantages and disadvantages in comparison. In doing so, I will attempt to show if (a) there are different stylometric signals, like authorial style, gender, and genre, (b) a particular feature selection favours the detection of such a style signal, (c) certain parameters, for example, distance measure, culling value, and z-score transformations, impact the signal's distinctiveness. In order to meet the requirement of literary contextualisation, methods and discussions will be deeply entrenched in literary theory, more specifically in Structuralism, and literary history. I will not be able to supply statistically sound null hypotheses for all these issues, as this is simply not my field of expertise. I will, however, try to summarise my research questions for each experiment in formalised null hypotheses to be able to better evaluate the outcome. In order to have as much control as possible over the data analysed, I have manually compiled a corpus, which comprises over 500 English literary prose texts from 1688 to 1914, thus covering both the *Long 18th century* and the *Long 19th century*.

All resources necessary to reproduce my analyses will be available online (`https://github.com/jbrottrager/stylcoR`). Due to copyright issues, I cannot share all the texts I have gathered, because some are not yet in the public domain. I will, however, provide access to the metadata table and to all frequency tables used to enable the

reproduction of and critical engagement with my results. All scripts will be provided in a way that facilitates re-use. In order to enable a user-friendly reproduction, the R-scripts will be combined into a package called `stylcoR`. This package covers all the required steps from pre-processing corpora and stylometric analysis to visualisations. Implementing all these steps in one environment and one package is not only more convenient for users, but also helped me to gain a deeper understanding of the processes applied and to decrease black boxing.

I will begin my thesis by supplying the necessary theoretical background for the methods and discussions. In the first section of this theoretical chapter, fundamental structuralist ideas and concepts will be introduced and discussed. This exploration will also give the opportunity to highlight why Structuralism is a particularly interesting approach in the context of CLS, but also to outline its shortcomings. The detailed treatment of Roman Jakobson's theories on language and literature, as well as his linguistic concept of distinctive features will form the core of this section. Following this, the first chapter's second section will then deal with the detailed examination and analysis of previous works from the field of CLS. I will especially focus on the way stylometric results have been interpreted and whether a consensus on the detection of certain stylometric signals can be determined. Additionally, a special focus will be laid on if and how a signal's detection has been linked to a theory from linguistics or literary studies. In the final section of the foundational chapter, I will discuss different approaches to corpus compilation and how corpora as the systems in which single texts are analysed influence results.

Following the theoretical discussion of corpus compilation in the first chapter, the second chapter will address the practical side of corpus selection and compilation. Here, I will outline how I have modified Mark Algee-Hewitt and Mark McGurl's approach to designing corpora (2015) to better suit the scale of my thesis and to make it generally more easily applicable. I will then go into detail on how I have created my corpus and how it was subsetted to produce smaller corpora for different steps of the analysis.

After describing the general data compilation and some preparatory steps, I will present three experiments that can be seen as different ways of approaching stylometry. The first of them will employ descriptive statistics to examine if a distinctive stylometric signal can be detected by creating subsets according to binary categories extracted from metadata. Additionally, it will explore to what extent such a possible signal is impacted by the composition and manipulation of features. For this purpose, a series of subsets is produced, using different feature selections, i.e. uni- or bigrams and different sizes of most frequent feature (MFF) lists, parameter settings, for example, distance measure, culling, and z-score manipulation, and categories for subsetting, like gender and generic

form. The second experiment will rely on supervised learning and will employ support vector machines (SVM) to classify individual texts into groups. Again, multiple iterations with varying feature selections, parameter settings and categories will indicate whether a stylometric signal can be detected and certain combinations benefit the identification. Finally, the third experiment will employ networks based on the distance values between individual texts to illustrate similarities and contrast between texts and clusters of texts. In order to make these models interpretable, they are filtered by applying a nearest neighbour or a percental cut-off method. The network visualisations will enable a close monitoring of changing clusters and will be used to examine whether certain categories are more likely to cause clusters when they are based on a particular setting.

Although each of these chapters will include a discussion of the results, a final comparison in the conclusion will offer the opportunity to compare and contrast the overall results. Moreover, there will be room for a broader interpretation of the results, summarising general tendencies and emphasising implications for similar projects.

Additional material, as, for instance, complementary visualisations, can be found in the appendix. The metadata and frequency tables, as well as the R-package `stylcoR`, can be found at `https://github.com/jbrottrager/stylcoR`; all the network visualisations used are available as zoomable interactive objects at `https://jbrottrager.github.io/jb/visualisations/`.

2 Theoretical Background and Previous Studies

In the following sections, I will discuss in some detail the theoretical basis of my approach, as well as previous works focusing on similar methods and/or subjects. The first section will give, as aforementioned, an overview of Structuralism and how it can be combined with CLS. The second section will provide a non-exhaustive review of previous stylometric contributions to show how their results have been interpreted. For better orientation, the projects will be divided up into four main research fields, namely authorship, gender, genre, and time. The final section will deal in a more theoretical sense with the issue of corpus selection and compilation; building on previous works, especially by Allgee-Hewitt and McGurl (2015).

2.1 Framework: Structuralism

Although Structuralism is a theoretical approach like many others, employing it in the analysis of literature—may it be in a more traditional or a CLS context—can sometimes, according to Jonathan Culler, be understood as ”a polemical gesture, a way of attracting attention and associating oneself with others whose work was of moment” (2004, 3). The underlying ideas of Structuralism are, however, not only ”extremely common [...] in mathematics, logic, physics, biology and all the social sciences” (Culler 2004, 3), but also lend themselves to modularised computational approaches.

Due to the manifold usage of *Structuralism*, many different definitions exist for the term and concept. Roland Barthes addresses this issue and points out that ”this word [Structuralism], most often imposed from outside, is today applied to projects that are very diverse, sometimes divergent and sometimes even antagonistic” (1997, 94). Thus, it seems vital to begin this theoretical section by choosing a suitable definition of the term

and concept for this thesis. In my understanding of Structuralism, I will follow Barthes, who pragmatically describes Structuralism as "a certain mode of analysis of cultural artefacts, insofar as this mode originates in the methods of contemporary linguistics" (1997, 95). This explicit link to linguistics is especially noteworthy, as Structuralism is a theory developed from linguistics, which is then applied to literary works, i.e. works of language (Barthes 1997, 95). Culler builds on this notion of a linguistic foundation, when he highlights the two main insights gained by a structuralist approach: "[F]irst, that social and cultural phenomena are not simply material objects or events but objects or events with meaning, hence signs; and second, that they do not have essence but are defined by a network of relations, both internal and external" (2004, 5).

These ideas clearly stem from Saussurean Structuralism. Social and cultural signs can be split up into their performance—i.e. the *signifiant*—and their attributed meaning—i.e. the *signifié*. Corresponding to Ferdinand de Saussure's description of the linguistic sign, social and cultural signs can only gain meaning through their inner contrast between *signifié* and *signifiant* and their outer contrast to other signs in a closed system:

> [D]'un côté, le concept nous apparait comme la contre-partie de l'image auditive dans l'intérieur du signe, et, de l'autre, ce signe lui-même, c'est-à-dire le rapport qui relie ses deux éléments, est aussi, et tout autant la contre-partie des autres signes de la langue. (2013, 248)[1]

What this also means is that no sign carries any meaning by itself; the entire system of signs is built on difference and difference only:

> Tout ce qui précède revient à dire que dans la langue il n'y a que des différences. Bien plus: une différence suppose en général des termes positifs entre lesquels elle s'établit; mais dans la langue il n'y a que des différences sans termes positifs. Qu'on prenne le signifié ou le signifiant, la langue ne comporte ni des idées ni des ons qui préexisteraient au système linguistique, mais seulement des différences conceptuelles et des différences phoniques issues de ce système. (2013, 258)[2]

From a methodological point of view, I hope it already becomes clearer why the combination of Structuralism and CLS approaches is so powerful: All methods which will be applied in later chapters represent stylometric analyses of the closed system of a

[1] [O]n the one hand the concept seems to be the counterpart of the sound-image, and on the other hand the sign itself is in turn the counterpart of the other signs of language. (Saussure, Baskin, et al. 2011, 114)

[2] Everything that has been said up to this point boils down to this: in language there are only differences. Even more important: a difference generally implies positive terms between which the difference is set up; but in language there are only differences without positive terms. Whether we take the signified or the signifier, language has neither ideas nor sounds that existed before the linguistic system, but only conceptual and phonic differences that have issued from the system. (Saussure, Baskin, et al. 2011, 121)

corpus. Individual texts are—in one way or another—compared to each other and are only attributed a specific meaning in the contexts of these comparisons. The attributed connotation can only hold true in the respective context of the analysis, i.e. the corpus that was used. In a different corpus, as will be seen in the experiments, the attributed meaning of each text can change significantly. To use Culler's words, each text "is itself structured and is defined by its place in the structure of the system" (2004, 5).

Even though the connection between structuralist linguistics and CLS has hopefully become clear, the question why a structuralist analysis of literary works can yield valuable insights is still unanswered. For this reason, I will now turn to Roman Jakobson's approach to Structuralism, which stresses the special role of literary texts for structuralist analysis. Even though I will heavily rely on Jakobson's connection between linguistics and literature and will employ his concept of distinctive features for the description of literary texts, I will not apply his notion of the poetic function in its entirety. This is primarily due to the fact that Jakobson works on a phonemic, grammatical, and morphological micro-level and examines structured sequences of sounds and syllables, which does not comply with my experimental design. Furthermore, he also introduces the idea of an "unbiased, attentive, exhaustive, total description of the selection, distribution and interrelation of diverse morphological classes and syntactic constructions" (1985a, 42). In this thesis, I will neither attempt an unbiased nor a total description of the chosen texts, as it is simply neither feasible nor—I would argue—possible.

	o	a	e	u	ə	i	l	ŋ	ʃ	ʧ	k	z	ʒ	g	m	f	p	v	b	n	s	θ	t	z	ð	d	h	#
1. Vocalic/Non-vocalic	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
2. Consonantal/Non-consonantal	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−
3. Compact/Diffuse	+	+	+	−	−	−		+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	
4. Grave/Acute	+	+	−	+	+	−									+	+	+	+	+	−	−	−	−	−	−	−		
5. Flat/Plain	+	−		+	−																							
6. Nasal/Oral								+	−	−	−	−	−	−	+	−	−	−	−	+	−	−	−	−	−	−		
7. Tense/Lax									+	+	+	−	−	−		+	+	−	−		+	+	+	−	−	−	+	−
8. Continuant/Interrupted									+	−	−	+	−	−		+	−	+	−		+	+	−	+	+	−		
9. Strident/Mellow										+	−		+	−							+	−		+	−			

Table 2.1: Distinctive Features in Received Pronunciation (Jakobson, Fant, and Halle 1963, 43)

As a linguist, Jakobson offers first and foremost an adaptation of de Saussure's linguistic Structuralism. Building on the Saussurean idea of binary oppositions (*signifié* and *signifiant*), he developed the theory of acoustic or distinctive features with Gunnar Fant and Morris Halle (1963). With these features, a given sound in a natural language can be described acoustically by attributing several binary values (vocalic/non-vocalic, for example). Table 2.1 displays the binary description of all vowels and consonants and their respective distinctive features in Received Pronunciation (RP). The chart, however, does not portray every possible combination of distinctive features, nor does it explicate all possible variations of a given sound. What it does is outlining the features which are necessary to distinguish the individual sounds and thus enable an inscription of meaning in them. Jakobson, Fant & Halle (1963, 9) elaborate on this point with the following example:

> There is a continuous variation in the shape of the lips from a close rounding to spreading and in the corresponding acoustic effects; but the linguistic opposition flat vs. plain (e.g. German Küste "shore" - Kiste "box") is a linguistic assignment of distinctive value to two distant lip positions and to their contrastive acoustical effects [...]

Distinctive features are thus not mere variations of a sound, but elements of sound productions which are recognisable for a receiver if they know the code system it stems from and if these elements have been transmitted accurately (Jakobson, Fant, and Halle 1963, 8).

In his literary analyses, or more precisely, in his analyses of the poetic function of texts, Jakobson employs the idea of binary oppositions to explain how literariness is achieved. To determine "[w]hat makes a verbal message a work of art" (1960, 350), a "focus on the message for its own sake" (Jakobson 1960, 356) must be detected. This detection is based on selection and combination, i.e. the two fundamental ways in which verbal behaviour can be arranged (Jakobson 1960, 358). In a non-poetic setting, word selection is based on the syntagmatic axis of lexical word choices. In sentences like, for example, "The hut is small." and "The house is small.", the selection of the noun is based on "equivalence, similarity and dissimilarity, synonymity and antonymity" (Jakobson 1960, 358). On the paradigmatic axis of combination, the notion of contiguity would normally determine the resulting word sequence. In a poetic context, however, "the principle of equivalence [is projected] from the axis of selection into the axis of combination" (Jakobson 1960, 358). In other words, each element in a literary text is placed in relation to all other elements of the sequence. By promoting "[e]quivalence [...] [as] the constitutive device of the sequence" (Jakobson 1960, 358), structures become more noticeable, or, as Culler phrases it in his discussion of Jakobson, "[p]atterns formed by the repetition of similar

items will be both more common and more noticeable in poetry than in other kinds of language" (2004, 66). Following Jakobson, one could thus argue that literary texts are not only valid subjects for a structuralist analysis, but even particularly suitable, as their structural settings are more easily detectable.

In his own analysis of poems, Jakobson offers another insight into the possibilities of Structuralism, when he bases his interpretation on counts of specific PoS and inflectional forms :

> The very selection of grammatical forms in the poem is striking. It contains forty-seven words, including a total of twenty-nine inflectional forms. Of the latter, fourteen, i.e., almost half, are pronouns, ten are verbs, and only five are nouns, moreover, nouns of an abstract, speculative character. In the entire work there is not a single adjective, whereas the number of adverbs is as high as ten. Pronouns—being thoroughly grammatical, purely relational words deprived of a properly lexical, material meaning—are clearly opposed to the remaining inflected parts of speech. (1985b)

This excerpt shows clearly how Jakobson's approach anticipated later developments in computational linguistics and CLS. By counting specific elements in a literary text, he indicates that structures cannot only be identified by examining sequences of text, but also by extracting frequencies of features. Comparing different texts based on their word counts (cf. Jakobson 1985b, 52-57) allows for a more distant structuralist perspective and helps to integrate Structuralism in a CLS paradigm.

The CLS point of view that I will take in the course of the experiments is in its foundations a structuralist one. From its considerations of a corpus as a system in which texts are set in opposition to each other, to the employment of distinctive features for the description of individual texts, and the usage of feature counts as the basis of the analyses, every step is rooted in a structuralist framework. Adhering to this theoretical system implies, however, a considerable drawback. A precise implementation of distinctive features means that only binary features can be chosen for the description of individual elements, may it be a phoneme or a literary text. This restriction is potent, as the description of more complex elements like literary texts inevitably causes simplification. For this reason, I have decided to follow the paradigm only to a certain extent and have introduced non-binary options in the analysis of genres (see Chapter 7). When trying to implement non-binary categories in computational contexts, it becomes more apparent than ever how many computational principles fundamentally build on binary oppositions, from binary code to logical operators. They thus facilitate the implementation of binary concepts and simultaneously hinder the usage of non-binary classes. Therefore, working with computational methods always

implies accepting the consequences of these formalisations and accounting for the caused limitations in the contextualisation of the results.

2.2 Methods: Stylometric Analyses and Their Interpretations

2.2.1 Authorship

Especially in earlier projects in the fields of stylometry and CLS, authorship attribution has been the prime goal. Frederick Mosteller and David L. Wallace employ a statistically informed approach in their seminal work on the *Federalist* papers (1963). Working with a Bayesian method for what they call ”discrimination” (Mosteller and Wallace 1963, 275) of authorship, they focus primarily on high-frequency function words, using a set of 165 unique words (1963, 281). By doing so, they are able to yield statistically significant results in the authorship discrimination for texts of disputed authorship. In their conclusion, they deduce that their study’s success can be traced back to the usage of function words as primary data source, as they ”appear to be a fertile source of discriminators” (1963, 306). Moreover, they stress that ”[c]ontextuality is a source of risk” (1963, 306). Contextuality occurs if a feature’s frequency varies considerably within the oeuvre of one single author and can therefore be assumed to be dependent on a text’s context. Their criteria for choosing features are thus that they are on the one hand frequent enough to be relevantly distributed in each text and on the other hand, that they do not—like, according to Mosteller and Wallace (1963, 306), pronouns and auxiliary verbs—display to much contextuality.

John Burrows, who developed the now ubiquitous distance measure *Burrows Delta*, has focused on authorship attribution in many of his works. With only few exceptions (2007), his projects, which range from verification tasks (2005) to methodological questions (2003, 2007), build primarily on a set of function words and/or high frequency words. Even though he references Mosteller and Wallace, he does not replicate their approach concerning contextuality. In general, however, Burrows’s word lists are comparable in length to Mosteller and Wallace’s, covering 60 to 289 individual words (Burrows 2002, 274, Burrows 2005, 443). Explaining his choice of features, Burrows distinguishes between strong features, i.e. semantically charged words typical for a specific author, and weak features, like function words:

> Strong features, perhaps, are easily recognized and modified by an author and just as easily adopted by disciples and imitators. At all events, a distinctive 'stylistic signature' is usually made up of many tiny strokes. (Burrows 2002, 268)

Burrows also points out that he is not trying to identify "unique authorial fingerprints (of whose very existence we do not yet have either proof or promise)" (2002, 268), but attempts to distinguish "the most likely candidates from a large group" (2002, 268) of possible authors. So even though Burrows is absolutely focused on authorship as a signal, his interpretation of probabilistic results is more nuanced. Nevertheless, this nuance seems to get lost when Burrows talks about the fact "that authors work at times in very uncharacteristic literary genres" (2002, 279), i.e. a possible genre signal, as the cause for poor results. Therefore, Burrows can be understood to be saying that, in a sense, style variations originating in genre are noise, obscuring the authorial signal.

A similar line of argumentation concerning the feature selection can be found in David L. Hoover's "Statistical Stylistics and Authorship Attribution" (2001, 422). In this contribution, he links the discriminatory quality of function words in authorship attribution to neurolinguistics, more specifically to findings by Angela Friederici in this field (1996, 178-179). These findings suggest that after the age of ten, speakers are able to process so-called closed-class words, as, for example, pronouns, prepositions, articles, and other function words, more quickly than open-class words, which include, for instance, nouns and verbs. Additionally, open- and closed-class words appear to be stored separately in the brain. Combined, these neurolinguistic insights are interpreted by Hoover as indicators of the possibility of an author's *wordprint*:

> Because of their high frequencies in the English language and their low semantic load, the most frequent function words have long been assumed to lie outside the conscious control of authors. If this is so, their frequencies should reflect deeply ingrained linguistic habits and should provide the analyst with what might be called an author's 'wordprint'. (Hoover 2001, 422)

Although many more studies, as, for example, Shlomo Argamon and Shlomo Levitan's "Measuring the Usefulness of Function Words for Authorship Attribution" (2005), Matthew Jockers's *Macroanalysis* (2013, 63-104), and Mike Kestemont's "Function Words in Authorship Attribution" (2014) at least partially support the proposition that authorship can be measured best when working with function words, there seems to be a considerable shift of attitude in later contributions. Beginning with Maciej Eder, who compares features and singles out differences for languages (2011), and Hugh Craig and Arthur F. Kinnley's *Shakespeare, Computers, and the Mystery of Authorship* (2012, 20,), who use lexical words in some of their experiments, a tendency to move towards more extensive feature

sets can be observed. Especially since Maciej Eder and Jan Rybicki's contributions on authorship attribution (Rybicki and Eder 2011, 2013, Eder 2016, Eder 2017), a practical exploration of larger feature sets begins: With this examination of a more diverse data set for authorship attribution, the focus shifts from the very restricted area of MFWs to what Burrows defines as "the large area between the extremes of ubiquity and rarity" (Burrows 2007, 27), i.e. features that are comparatively frequent but not necessarily part of the top 100 MFWs.

In the context of feature selection and interpretation, Eder and Rybicki's "Deeper Delta across genres and languages" (2011) seems especially interesting, as it tests a myriad of different feature sets and their influence on correct authorship attribution. Feature sets of several sizes are pushed through the entire feature list—i.e. starting at position 1, then at position 50, then at position 100—and the respective results are then compared and analysed. For an English-language corpus, the classification results are generally exceptionally good. The accuracy of these results, is, however, not the most significant information gained from these examinations: As Figure 2.1 shows, the attributive success increases with the size of the feature vector (see the dark red areas along the x-axis).

With these results, Eder and Rybicki (2011) challenge the notion suggested by more traditional approaches that authorship attribution is most successful when relying on a very limited number of MFWs. But even though they go into a very specific and morphologically informed discussion on why the attribution works best for an English-language corpus (2011, 319-321), no further explanation for the link between better classification results and more extensive feature sets is offered.

Following Eder and Rybicki (2011), many scholars have investigated authorship based on feature sets of up to 10,000 MFWs. Similar to Eder and Rybicki's comparative methodological approach, Stefan Evert et al. (2015), Evert et al. (2017), and Andreas Büttner et al. (2017) compare the accuracy of combinations of feature sets and distance measures. The discussion of the attribution success is, however, contextualised to an even lesser degree of detail than in Eder and Rybicki's contribution: Any attributive success—may it be for a feature set of 50 or 5,000 MFWs—is linked back to the authorship signal. Unfortunately, this link is established without any explanation drawn from linguistic or literary theory. In other words, the validation of the approach and method is solely based on the seemingly correct attributions.

A subtler approach is chosen by Stefan Schöberlein (2016) and Michael Oakes (2018), who both vary their feature sets in different experimental designs. Depending on the respective setting, they choose different sizes of feature sets and compare the results. However, there is again a lack of theoretical background in the discussions of the results,

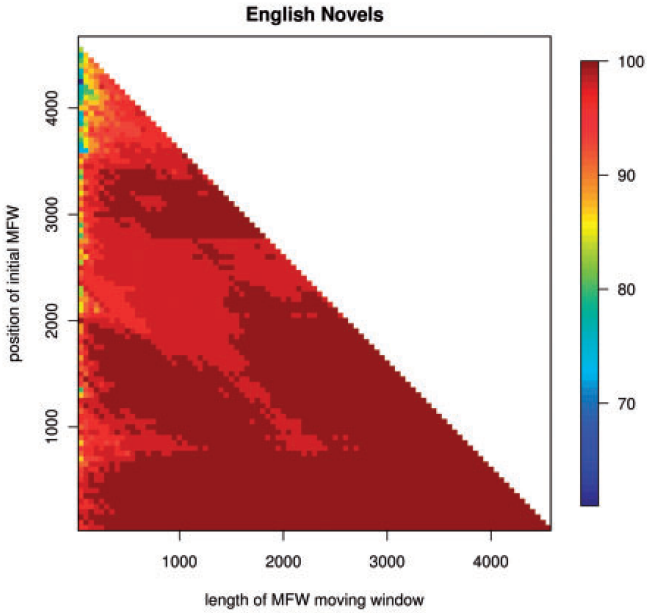

Figure 2.1: Accuracy of Authorship Attribution in an English-language Corpus for Different Feature Sets (Rybicki and Eder 2011, 317)

as the validity of the feature selection is linked only to the rate of correct attributions. Interestingly, Oakes identifies the difference between the very broad genres of fiction and non-fiction as discriminatory signals (2018, 646). By doing so, he addresses the fact that feature sets do not measure one specific characteristic of a text and that this must be acknowledged in a more conscious interpretation.

2.2.2 Gender

A characteristic that is often analysed in similar terms as a text's author is the authorial gender. Based on the partially disputed theories on specific features of female language by Robin Lakoff (1973, 2004) and more recent contributions on gender-specific language usage by, for example, Janet Holmes and Miriam Meyerhoff (2003), scholars try to differentiate texts according to their author's gender. Similar to the detection of authorship, the authorial gender has been examined on the basis of many different feature sets.

A major trend in the differentiation of authorial gender is the comparison of the usage of function words and/or high frequency words. Argamon et al. (2003) and Jockers (2013, 63-104) rely heavily on these features for the attribution of gender. They do, however, also use the same feature set for the analysis of authorship (see Argamon and Levitan 2005, Jockers 2013, 63-104). While Jockers does not vary his feature selection in any way, Argamon et al. apply additional weights to contrast the use of the chosen function words in texts by female and male authors. Koppel et al. (2002) introduce complexity on a different level, when they do not only consider function words, but also counts for specific PoS tags and sequences. By doing so, they are able to monitor syntactical and grammatical characteristics of gendered language, which represent, according to Lakoff (1973, 2004), core differences in the language use of female and male speakers.

Others have chosen a more specific approach to gender disambiguation: Rybicki (2016) and Sean Weidman and James O'Sullivan (2018) employ—inter alia—a so-called zeta procedure to extract features which are especially common in one group of texts (e.g. texts by female authors) and uncommon in another group of texts (e.g. texts by male authors). Building on these lists of discriminatory words, they perform attributions that only rely on these predefined features. As an explanation of their discriminatory quality, both Rybicki and Weideman and O'Sullivan point out that a close reading of these lists reveals gender stereotypical distinctions of the female and the male sphere, which align themselves with linguistic theories proposed by, for example, James Pennebaker (2013).[3]

But why are these lists so full of words linked to gender stereotypes and what are the consequences of using these lists for a attributive analysis? Logically, by extracting words which are particularly common in all individual texts belonging to a certain group, but particularly rare in texts from another group, a corpus's entire lexicon is reduced to extreme outliers. When these outliers are used to define what typical female and male style is, the resulting attributions will follow these polarising definitions. This means that if a text's authorial gender is identified as female—may it be in a classification or by clustering with other texts by women—it can only be deduced that its counts for words representing extreme gendered style are more similar to other texts by women. In other words, what is measured in such a procedure is not necessarily a specific authorial gender signal, but whether a text gravitates towards one extreme of gendered language or another.

An additional variant of the use of predefined lists is employed by Ted Underwood (2018),

[3] Pennebaker claims, for example, that women make more use of first person singular, cognitive and social words, personal pronouns, verbs, negative emotions, negations, certainty words, and hedge phrases and men lean towards "big words", nouns, prepositions, numbers, and swear words (2013, 40-43).

when he algorithmically defines content words which are used in proximity to a story's characters as features. By doing so, the strategy of gender representation is used as a proxy for the author's gender. Similar to the approach described above, it is questionable whether this proxy in fact captures female and male style, especially when the results (see Underwood, Bamman, and Lee 2018, Figure 10) are in most parameter settings—if at all—only slightly better than chance.

2.2.3 Genre

To a lesser extent, stylometric analyses have been focused on the attribution of genres. Jockers, who, as aforementioned, inspects different characteristics with the same set of function words and high frequency features (Jockers 2013, 63-104), is one of the first to test and discuss under which circumstances attributions are caused by a genre signal. In his analyses, the correct genre is attributed for 67 percent of all test data, representing an attribution that is eight times better than chance (Jockers 2013, 81).

Despite this comparatively high accuracy, others have challenged Jockers's usage of high frequency words for the disambiguation of genres. Christof Schöch, for example, claims that "genre, most likely because it is strongly related to themes, is more likely to show up in parts of the wordlist beyond the function words" (2012). For his specific case study, a contrastive examination of Classical French plays, the best results were yielded around 850 MFWs. In a more recent project, Schöch even uses feature sets with up to 1,000 MFWs (2014). However, Schöch asserts that, at least in his results, it is not possible to distinguish one textual characteristic from another, as "the author always somehow shows up along with genre" (Schöch 2012).

A similar size of feature sets is employed by Underwood (2019), who uses between 1,100 and 2,200 features. Generally, he relies on a MFW approach, even though he points out that there are some modifications he employs, like summarising personal names, place names, and days of the week in one variable or including macro-level features like average word and sentence length (2019, 196). Most significantly, Underwood goes into detail to underline that feature selection might not be influential at all:

> But generally, I try to avoid spending a great deal of effort on feature selection and engineering. For one thing, it doesn't help. I have spent weeks designing systems that assess rhyme and meter, measure conditional entropy in fiction, or count phrases longer than a single word. But these features almost always duplicate information that was already latent in word counts. (2019, 196)

It is interesting that Underwood comments on the qualitative aspect of feature selection, but does not elucidate his quantitative feature selection. It might be the case that in the context of Underwood's research project, more complexly generated features, as, for example, rhyme and metre, do not contribute in any way to more reliable and insightful results. Others, like Douglas Biber (1988, 2009), have shown quite convincingly that, at least for linguistic genres, the usage of PoS tags and sequences does lead to relevant differentiations of genres. The assumption that different kinds of features do not lead to significantly different results might therefore be misleading. Additionally, Underwood does vary the size of his feature set considerably. This can be understood as an implicit acknowledgement that at least the feature set's scope influences an analysis's outcome.

2.2.4 Time

In his discussion of detectable gendered groups, Rybicki makes a valuable point about an additional category, which should be examined in more detail:

> I wonder if time, perhaps treated more generally, is not much, much more important. There is, it must be said, an interesting dualism about the chronological signal in literary language, since it concerns single-lifespan and single-author collections of texts as well as large and long-span multi-author corpora, and both phenomena cannot be blamed on the same mechanism of linguistic change. (2016, 759)

Temporal characteristics have been discussed to some extent by Jockers, when he examines the way texts from different decades are classified (2013, 63-104). For this classification, he again uses his predefined set of function words and high frequency features, achieving slightly more than 50 percent accuracy (2013, 81).

Other scholars, like Franco Moretti (2005) and Andrew Piper (2018), deal extensively with time and temporal qualities of literary works, but tend to treat time as an independent rather than a dependent variable, examining, for example, how genres change over time. This means that if such relationships are plotted, the temporal categories are displayed on the x-axis, the respective dependent variable on the y-axis. By doing so, possible changes in style are detected in relation to the dependent variable, and not necessarily in relation to temporal qualities.

2.3 Context: Corpus Selection and Compilation

Corpora are the essential data sources in stylometry and CLS. As a data basis, their influence on results cannot be underestimated. In recent years, several different strategies to corpus selection and compilation have been applied; some exemplary cases will be discussed in the following. Each of these strategies has a slightly different motivation: Some corpora are created as means to an end and are compiled with a very specific research question in mind. Others try to mirror more generally the literary field of a specific time span and therefore aim for representativeness. What all these approaches have in common is that on the one hand, they need to lead to corpora which produce reliable results and on the other, they have to be practicable. Finding the middle ground between representativeness and realisability is of utmost importance: No corpus can comprise all literary works of a given era or thematic focus, but they need to cover enough of them to be able to make sound judgements about the respective part of literary history.

But how much is enough? Many would probably argue that more data leads to better results; the argument being that big data eventually covers all varieties of a population. This might seem logical, but especially for literary corpora, the question of the quality of selection has to be set before the question of quantity. This has several reasons: First, literary corpora are often too small to be defined as big data. Collections of several hundred works will never be able to summarise the vast amount of texts that have been published. Second, availability plays an enormous role in the compilation of corpora. Texts that are already digitised and accessible online can be incorporated more easily in a corpus. Their availability is, however, often linked to their standing in the canon. Thus, using all available texts does not mean that all literary history can be covered and examined, but only that a predefined sub-group of literary texts is explored. This inherent bias needs to be balanced out to yield dependable results. Third, text data is prone to errors and therefore requires clean-up processes. Although there is some research suggesting that the influence of flawed textual data and errors is comparatively small (Eder 2013), this issue has yet to be explored in more detail, especially as it has only been tested in the context of authorship attribution. Until decisive results are provided, it is advisable to correct textual errors in corpus texts carefully and consistently.

The probably most convenient way to tackle the issue of corpus compilation is to use an already existing corpus for an analysis. They often have a very specific thematic focus and are generally curated on a high scientific level. In his article on stylistic gender differences (2016), Rybicki uses such a corpus, namely the Chawton House Corpus (*Novels Online* 2016), which collects little-known novels by women from 1723-1830. Similarly highly

specific sources can be found in the Chadwyck-Healey database collection (*Chadwyck-Healey Databases* n.d.), which provides access to corpora from *African-American Poetry* to *Early English Books Online* (EEBO). The advantages of using pre-existing academic corpora are obvious: First, they are ready to use and often downloadable in different formats. Second, as they are compiled by academic institutions, they can be assumed to be based on scientific grounds and thus representative for their chosen subject. Third, as scientific staff was involved in their creation, the texts themselves are more reliable and less prone to errors. Additionally, commentaries are often supplied to give insight into editorial decisions and possible alterations. As there are, however, comparatively few of these corpora and many of them are only available for licensed users, not too many research projects use them.

A very different strategy that has become increasingly popular with growing free online plain text archives consists of automatically scraping or manually copying texts from these platforms and thus creating corpora for specific research purposes. Examples of this approach are Evert et al. (2015) and Jannidis et al. (2015), who use three language-specific corpora for English, French, and German literature. They compile there text collections with the help of Project Gutenberg, Ebooks libres et gratuits, and TextGrid, respectively. Due to the research focus on authorship attribution, the question of representativeness and canon bias is neglected in both papers. More generally, no additional information about the texts used is provided, except for the respective covered time frames (Evert, Proisl, Vitt, et al. 2015, 81-82, Jannidis et al. 2015 1-2) and the fact that each single author is represented by three texts. For the methodological framework of these papers, these shortcomings might be acceptable, as no interpretations about any specific authors or texts are offered.

Another example of using online resources is Underwood's latest project *Distant Horizons* for which he compiled a corpus using the platform HathiTrust[4] (2019). There are, however, some major difference to the strategies described above: Underwood's corpus does not consists of hundreds, but thousands of individual texts and thus can actually be defined as a big data collection (although he sometimes filters the main corpus of 93,960 volumes and then works with a resulting sub-corpus of 347 volumes, for example). Despite this enormous size, he spends some considerable time to outline the corpus's general structure (2019, 173-184) and describes how automatic processes were applied to generate metadata on the one side and to clean up the texts, on the other. The texts' spelling was, for instance, "[w]henever possible" (2019, 182) normalised to the modern British variant. More importantly, Underwood deals very consciously with the shortcomings

[4]HathiTrust can only be used in its full extent by affiliates of a contributing organisation.

of his corpus when he highlights that there are certainly many errors left in the data (2019, 182) and that the data quality is not perfect, but "good enough to answer [...] broad questions" (2019, 184). Furthermore, he deals in some detail with questions of corpus bias and text availability in online archives (2019, 173-181). Underwood suggests two ways to check such a corpus bias: First, he compares results yielded from his main corpus with those from an alternative corpus, namely the Chicago Novel Corpus. This corpus is considerably smaller than his main corpus and was compiled manually. If the results are similar—as they are—it can be assumed that the automatically collected and cleaned corpus is as representative as the more carefully compiled one (2019, 131-133). Second, and, according to Underwood, more importantly, he employs re-sampling and testing to account for uncertainty rates (2019, 180-181). All this shows that even though Underwood might not be working with the perfect data, he is aware of its limitations and explicitly elaborates on them (2019, 177-181).

Similar to Underwood, Algee-Hewitt and McGurl (2015) call for a more conscious occupation with corpus selection and compilation in their planning of a representative corpus of 20th century novels. In contrast to Underwood, they propose a more time-intensive strategy that involves a fundamental discussion of the relationship between the canon and a corpus. Thus, their approach does not, like Underwood's, build on sampling and re-sampling, but on a careful selection of corpus texts to achieve reliable results. They propose working with a couple of lists which rank works of literature according to their quality, popularity, or relevance. The first aspect is covered by lists featuring the supposedly best novels of the 20th century, which focus on expert opinions. For the second facet, contemporary and present-day audiences' opinions are included by taking into account lists of best novels compiled by readers and lists of best-sellers. In order to cover all relevant texts, they additionally use lists produced by academics from fields like Feminist Literary Studies, Post-Colonial Studies, and Multi-Ethnic Literary Studies. By doing so, Algee-Hewitt and McGurl attempt to counterbalance possible data biases to create a corpus which does not only mirror the canon, but also includes non-canonical works and light fiction. As their contribution is first and foremost an outline of their selection processes, they do not expand on the details of the actual corpus compilation. Since they plan to create a corpus similar to those provided by the Chadwyck-Healey collection (2015, 1), it can be, however, assumed that the digital texts will be created according to similarly high academic standards.

3 Corpus Preparation and Subsetting

Considering the discussions of previously applied strategies for corpus compilation, there are a few lessons to be learned for the compilation of my own corpus. Unfortunately, there is not yet a ready-to-use corpus for the time span of my analysis, which means that I have to compile one myself. For this compilation, I will, due to the scope and focus of my project, lean heavily on Algee-Hewitt and McGurl's approach to balanced corpus selection (2015). The steps proposed by them are very specific for their research subject of 20th century literature; in order to be able to apply them, they need to be systematically adapted. There are, for example, no lists of most significant literary works from 1688 to 1914, even for partial epochs of this time span, there are hardly any undisputed lists available. The same is true for lists of most popular works, as sales figures are only partially available. Additionally, due to the scope of this paper, it is not feasible to distribute questionnaires to experts in specific literary fields to create a list of non-canonised relevant works. As an alternative to all these lists, it seems most practicable and thorough to search secondary sources for mentions of primary texts and thus put together a corpus list covering all pertinent texts.

My choice of secondary sources reflects the different levels of canonicity covered by Algee-Hewitt and McGurl: They incorporate the very narrow and restrictive notion of canon, but also a broader academic canon and non-canonised literature by marginalised authors, i.e. women and writers from the geographical and linguistic periphery, as well as generally devalued genres, i.e. light and popular fiction. To include the higher ranks of canonicity, in other words texts that are considered to be crucial for a given time period and thus keystones of literary history, the *Norton Anthology of English Literature* (Greenblatt and Abrams 2006) and the literary historical source *English Literature in Context* (Poplawski 2008) were used. For a broader definition of the canon, I looked through several companions to genres and epochs (Caserio 2009, Curran 2010, David 2012, Herman 2007, Marshall 2007, Maxwell and Trumpener 2008, Shattock 2010). Additionally, I included sources in my research that explicitly deal with literature by

women (Ingrassia 2015, Looser 2015, Peterson 2015) and literature stemming from non-standard language traditions and more peripheral areas of the British Isles (Carruthers and McIlvanney 2012, Foster 2006). Most of these publications already address to some extent genres and forms of literature that were not or are still not regarded as high literature. To ensure that these forms of light and popular literature were sufficiently represented in my corpus, I also used specific chapters on popular fiction (Flint 2012).

Not all literary works mentioned in the sources eventually ended up in the corpus. This is because I introduced some more restrictions regarding the corpus's general design. In its essence, it is an English-language corpus—texts written primarily in other languages, like Gaelic, were not included. This does not mean that I do not consider Gaelic literature to be part of British literary history, but only that it is methodologically problematic to introduce a text in a different language in an otherwise quite homogeneous corpus. Concerning the broad generic form of texts, I have only included works which can be defined as fictional prose texts. By this definition, I excluded memoirs, biographies and autobiographies. Texts belonging to the overall thematic genre of life-writing were only added if they showed a significant degree of fictionalisation, as, for example, Thomas Quincey's *Confessions of an English Opium Eater* (1821).

The choice of the designated time span for the corpus might seem arbitrary at first, but is motivated by two reasons. First, literary corpora are often designed to represent a century in literary history (see, for example, the Chadwyck-Healey database of *Nineteenth-Century Fiction* and Gale's *Eighteenth Century Collections Online database*). These temporal restrictions to a century might be practicable, but also support a more fractioned understanding of historical developments. A corpus covering a larger period of time can be assumed to be able to better illustrate such broader trends. Second, literary history can never be interpreted out of its general historical context. Thus, a suitable time frame should consider more general events and periods in history. The time span chosen comprises the *Long 18th Century* and the *Long 19th Century*. The *Long 18th Century* covers the years from the Glorious Revolution in 1688 to the Battle of Waterloo in 1815; the *Long 19th Century* the years from the beginning of the French Revolution in 1789 to the beginning of World War I in 1914. Choosing a temporal frame that is not specifically linked to a national (literary) history will hopefully enable a comprehensive investigation of literary texts and does also provide the prerequisites for future usages in the context of a bi- and multi-language corpus project.

A text that meets all the aforementioned criteria, i.e. an English fictional text published between 1688 and 1914 which is mentioned in one of the chosen secondary sources, was added to the corpus list. This list eventually built the basis for the actual corpus

compilation. For the compilation, I started by searching for each text in openly available full-text archives, like Project Gutenberg. Although these platforms provide access to a vast amount of texts, not all texts on the corpus list were available. For these cases, I tried to find alternative sources, as, for instance, PDF editions and scans. If neither a plaintext nor a PDF version of the text was accessible and if the individual text seemed significant for the balance of the corpus—i.e. if its relevance was particularly stressed in the secondary sources or if no other text by the author was already in the corpus—I retro-digitised[1] it. As retro-digitisation is a time-intensive process, these restrictions were necessary to keep the corpus compilation practicable. After collecting both texts and corresponding metadata, I cleaned up all texts individually. This pre-processing included omissions of metadata and paratexts, as well as normalisations of common digitisation errors and formatting irregularities.

3.1 Metadata and Distinctive Features

The metadata table, which provides several data points about each work, is not only helpful for gaining some overview over the collected data, but is also essential for many steps in the analysis. Besides information about the collection process—for example where and in which format the text was found—the table offers information on the author (name, gender, and nationality), the text's publication (earliest publication data, possible serialisation, and media outlet), and generic classifications. There are of course many more possible metadata categories such a table might include and this selection covers only very broad options of describing a literary text. The categories were primarily chosen because they are established classes in CLS projects and generally broadly used in literary studies.

In order to be able to use them purposefully, these categories had to be normalised and formalised. Following the model of Jakobson's distinctive features, I implemented binary classes to capture the aforementioned metadata points partially or in their entirety. Exceptions for this paradigm are NA options (meaning not applicable), which were used in rare cases to indicate that a piece of information is missing, and the class of specific genres, which were not systematised in a binary scheme. The same applies to the category of authorship, which is was not separately introduced to the metadata table, but which is featured in each filename. Due to the specificity of some methods, however, NA cases as

[1] Retro-digitisation is a process involving the scanning of a physical source and using Optical Character Recognition (OCR) and manual proof-reading to create a digital text version.

well as non-binary classes could not be considered for the first two experiments and will only come into play for the last one.

gender	nationality	threshold_1815	novel	epistolary	genre
female	English	before_1815	novel	epistolary	adventure
male	non_English	after_1815	non_novel	non_epistolary	bildungsroman
NA	NA				children_juvenile
					comic
					detective
					domestic
					fantasy
					historical
					industrial
					life_writing
					novel_of_manners
					philosophical
					picaresque_satire
					romance
					science_fiction
					sensation_gothic_mystery
					sentimental
					social
					NA

Table 3.1: Normalised and Formalised (Distinctive) Features

Table 3.1 illustrates all classes used in the analyses with their respective possible options. The first of these categories, gender, builds heavily on the social performance of gender, and is, in order to correspond to a binary scheme, restricted to the options of female and male. This choice can be claimed to reinforce gender stereotypes, as men and women are presented as opposites. In a way, this already shows the problematic tendencies inferred from implementing a structuralist approach to literature: There is not much room for nuance. A structuralist analysis of male and female style will thus always yield results which plot these groups against each other, simply because no additional options are available. As a result, a model based on these dichotomies will only be able to prove similarities caused by gender with the condition of assuming that the categories of male and female are invariable. This presumption of stable gender identities can, however, be dismissed in reference to Judith Butler's theory of gender performativitiy (2006) and has already been practically rebutted by Underwood (2018), who shows how the reliability

of gender detection declines over time. Keeping these restrictions in mind, the variable gender will, nonetheless, be used to illustrate broad stylistic differences between authors along the lines of social gender. The NA option is not used to diversify this distinction, but only for cases of unknown authorship.

Similar constraints apply to the second category of nationality. Here, I have decided to place the more dominant centre, England, against the more marginalised peripheries, Scotland and Ireland, and in rare cases, other nationalities (e.g. Nigerian-British, Australian-British, and Franco-English). The NA option is again reserved for anonymous authors or for authors for whom too little biographical information is available to determine their nationality. As an author's nationality can in most cases be linked to a certain language variety, this category can also be seen as a formalisation of the dichotomy between a more prestigious standard variant of English and less respected regionally influenced dialects. This is, again, a stark simplification of linguistic diversity and code-switching processes. Generally, there are many local English dialectal variations that are not seen as prestigious, and it can be assumed that the majority of non-English authors switch to a standard form of English for their writing.

Even more simplifications were introduced concerning the temporal classification of texts. In order to be able to use publication dates in a binary category, they had to be summarised in two very general groups, i.e. before and after a certain date. The definition of a temporal breakpoint is always arbitrary, as no shift in style ends in one year and begins in the next. Choosing 1815 as this cut-off point takes the historical context into account, as it represent the end of the *Long 18th Century*. Choosing a historically informed threshold for splitting up literary works into groups might prove to be a suitable approach to examine general temporal style patterns.

The generic classification of literary texts is often problematic even in a traditional literary studies setting. This is due to the fact that the term *genre* is used in various ways, applying to several structural and thematic levels. For instance, it is used to distinguish the main genres of prose, poetry, and drama but also for describing primary and secondary subgenres within these classes. Furthermore, the concept of genre is also connected to structural characteristics, when, for instance, novels are defined as epistolary novels or shorter prose as short stories. For these reasons, I have split up generic categorisations into three metadata points, which help to distinguish the different concepts of genre. First, the category of *novel* declares whether a given text is a novel or not. In other words, this category attempt to formalise the concept of generic form and separates novels from all other forms of prose (e.g. tales, short stories, and fables). Second, the metadata point *epistolary* helps to distinguish texts on a level of structural features. It can be assumed

that a text consisting of letters, which therefore has very clear structural features, will be analysed differently than a text without such striking particularities. Finally, the category *genre* offers the opportunity of a more fine-grained differentiation of sub-genres. With a `NA` option for unclear cases, there are 19 possible classes of sub-genres, which obviously does not cover all the potential genre options. Nevertheless, the selection offers a broad variety of labels and covers or summarises most of the genres mentioned in the secondary sources. With its 19 possible classes, this category is, as aforementioned, not adhering to the general dichotomous scheme of distinctive features. Even though this hinders the category's implementation for some experiments, more detailed insights into the networks can be gained, which is why this decision seems justifiable.

3.2 Corpora

The corpus compilation described above resulted in a corpus comprising 561 prose texts by 152 authors. Inspired by Underwood's idea of a smaller control corpus and based on the theory that a corpus as a structural context has an immense influence on possible outcomes, I have decided to subset the corpus into two smaller corpora. Additionally, the networks based on the smaller corpora are easier to interpret. These smaller collections, the midi- and mini-corpus, can be used to trace developments through contextual changes: The mini-corpus consists of 111 texts and was arranged in a way so that each text has several options for building relationships based on similarity. Of the 73 authors featured in the corpus, more than half (41) are represented with two texts. This means that if a given parameter setting favours the detectability of similarity based on authorship, it can be assumed that these 82 texts will probably appear in pairs. On the other hand, each of these texts belongs to groups based on other factors, like genre and time. If another parameter setting thus benefits the detection of other signals, similarity structures will hopefully be formed based on these grounds. The midi-corpus consists of all texts featured in the mini-corpus and of 72 additional texts. These texts add a considerable amount of complexity, as the number of individual authors is enhanced by more than 20 percent. Each of the 29 added authors is represented by two texts, all other texts are supplementary texts by already included authors. Again, this offers a multitude of opportunities for signals to be detected and the additional texts also help to confirm or refute findings of the mini-corpus.

As mentioned in the section on Structuralism, corpora are closed systems, which determine the possible interpretations for each texts featured in them. If the corpora are, however,

similarly structured, it can be assumed that these interpretations correspond to a certain degree. In order to be able to make this connection, the proportions of groups should be similar in all corpora. The following figures (Figure 3.1 - Figure 3.5) illustrate the relative category proportions of all three corpora. Regarding the category of gender, 28.52 percent of texts in the main corpus were written by women, in the midi- and mini-corpus, this ratio is considerable higher (39.01% and 34.23%, respectively). The rise is primarily due to the fact that I have tried to include as many genres as possible in the mini- and midi-corpus. Works belonging to some of these genres, like, for instance, children's literature, are—at least in the corpus—mainly attributed to women. In contrast to this, the ratio of English and non-English authors and the proportions of epistolary and non-epistolary works are very similar in all corpora. The share of English authors ranges from 58.11 percent (corpus) and 61.26 percent (mini-corpus) to 62.63 percent (midi-corpus); that of epistolary texts from 1.96 percent (corpus) and 2.70 percent (mini-corpus) to 3.3 percent (midi-corpus).

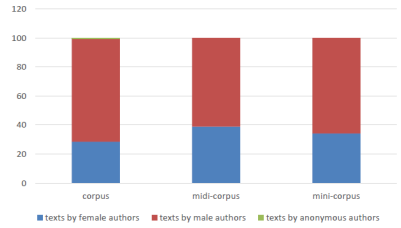

Figure 3.1: Relative Gender Proportions in All Corpora

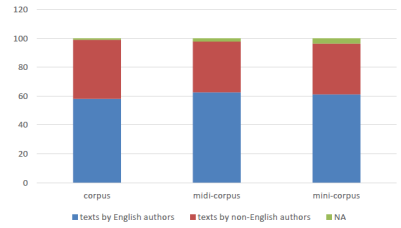

Figure 3.2: Relative Nationality Proportions in All Corpora

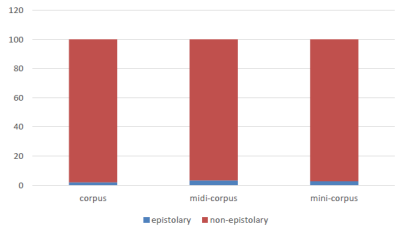

Figure 3.3: Relative Proportions of Epistolary Works in All Corpora

Similar to the gender proportions, the ratio of texts published before and after 1815 is not stable due to reasons of genre inclusion. Some genres, like sentimental novels, were more popular in the 18th century, which is why the relative proportion of text published

before 1815 is raised to include some of these texts. As many sub-genre conventions concern novels, the relative proportion of novels is also considerably higher in both midi- and mini-corpus: In the main corpus, 64.88 percent of all texts can be classified as novels, whereas this number rises to over 80 percent in the midi-corpus and to over 75 percent in the mini-corpus.

Figure 3.4: Relative Temporal Proportions in All Corpora

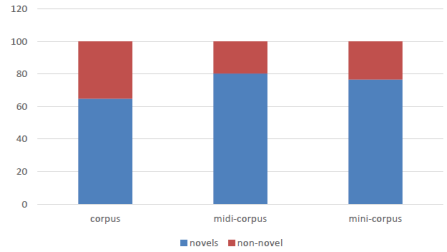

Figure 3.5: Relative Genre Proportions in All Corpora

Despite these inconsistencies, it can be presumed that the approach of detecting patterns of similarities throughout all corpora is still valid, as the discussed variations of proportions are considerable, but still stay within reason and do not, for instance, turn entirely. Additionally, it seems more important to include as many different genres as possible in both sub-corpora in order to be able to compare the findings productively.

4 Preparatory Work

All methods used in the experiments (chapter 5, chapter 6, chapter 7) are fundamentally based in one way or another on the results of several preparatory steps. These steps basically cover the assembly of frequency tables and the subsequent calculation of distance tables (see functions `createFreqDist` and `createDistanceTable`). As my experimental set-up consists, similar to Eder and Rybicki (2013), of iterations of parameter manipulations, several frequency tables with different features (uni- and bigrams) are assembled and filtered according to values chosen for the culling and the MFF cut-off. In sum, 120 iterations were conducted for each corpus, resulting in 360 sets of results. In the following, I will discuss the feature and parameter options I have used for my analysis and will provide some technical details of their implementations. This description will not offer a comprehensive overview of all functions and their implementations that are available in the R-package `stylcoR`. If you are looking for more practical and thorough information, please consult the package's documentation available at `https://github.com/jbrottrager/stylcoR`. Although I have used a myriad of sources, guidelines, and StackOverflow posts in the creation of this package, the general outline of the package relies heavily on the excellently structured and well-documented R-package `stylo` (Eder, Rybicki, Kestemont, and Pielström 2019).

The script I have used for calling the functions for all analyses can be found in the appendix (section 9.1). Running the whole script, especially for larger corpora and when using bigrams, might take a considerable time. For this reason, several update messages and progress bars have been introduced into some functions, which should provide the user some overview of the progress.

4.1 Features and Parameter Settings

The first setting which comes into effect is the choice between uni- and bigram, as the feature selection is the most fundamental level of my implementation. Having an accumulative purpose, the main function `createFreqDist` calls six modularised functions in the correct order and ensures the passing of correct parameters and values. In the first of these modularised functions, `createCorpus`, the input texts are processed into a `quanteda` corpus with the help of functionalities provided by the `quanteda` (Benoit, Watanabe, et al. 2018) and the `readtext` package (Benoit and Obeng 2019). This step includes the removal of white spaces and punctuation marks and the conversion of all letters into lower-case by default, which can be adapted by changing the corresponding boolean parameters in `createFreqDist`. The tokens of this created corpus are then, if applicable, turned into bigrams (`ngram`), again using a function from the `quanteda` package. On this basis, relative frequency distributions of all features in all texts are calculated. In a second step, the uni- or bigrams are also used to compile a sorted list of the maximal 1,000,000 MFFs in the corpus.[1] Using this feature list, the frequency distributions are aligned and combined in a matrix comprising all texts as rows and all features as columns.

This is when the next parameter, the culling value, comes into play: Features which do not appear in a specific number of texts are omitted. If, for example, the culling value is set to be 20 percent, a feature has to occur in at least 20 percent of all texts to qualify. The function `performCulling` is the first of seven possible functions called by the second accumulative function, `createDistanceTable`. Following the culling, the now limited matrix of feature frequencies is limited to the designated MFF cut-off (`cutMFF`). The resulting frequency table is saved as a csv-file and then used to calculate the z-scores for each feature (`calculateZscores`). Z-score calculation is a statistical method that enables taking into account the mean value of a feature; a z-score itself represents the number of standard deviation a given score is removed from this mean value. If the parameter for the z-score transformation is not ”none”, but either ”normalise” or ”ternarise”, the respective transformation is applied.

The two z-score transformations implemented in `stylcoR` are fundamentally different

[1] I have set this limit to enhance the package's efficiency, although this efficiency comes with a cost: In one parameter combination (corpus, unigram, 3,000MFF, 80% culling), the limit results in a smaller number of features, as only 1,979 tokens in the 1,000,000 list of tokens occur regularly enough to survive the culling. For future implementation, other work-around have thus to be found to ensure both efficiency and completeness of results (see chapter 8).

in their effect. The normalisation can be seen as another step in regularising the data, as the z-scores of each feature are normalised using the 2-norm. Normalising the data in this way should further minimise the influence of text length on the results. Ternarisation, on the other hand, is a method to amplify differences between groups of texts and to thus create stronger contrast between them. This is achieved by dividing z-scores into three groups: Z-score values below -0.43 are replaced by -1, those between -0.43 and 0.43 by 0, and those above 0.43 by 1. Assuming a normal distribution of values, this segmentation leads to roughly even groups. The resulting z-score tables, with or without transformation, are again saved as csv-files.

The final step of the fundamental data generation is the compilation of distance tables. These tables measure the distances from each text to every other text and can thus also be described as adjacency matrices. There are many different measures which can be applied in the calculation of the distances; I have decided to implement the most established ones, i.e. Burrows Delta and Cosine Delta. Each of the distance measures is implemented in an individual function, i.e. `caculateBurrowsDelta` and `calculateCosineDelta`.

5 Experiment 1: Descriptive Statistics

The first experiment I conducted tries to approach the issue of stylistic signals and their detection by examining subsets of the corpus with the help of descriptive statistics. Step by step, the distance tables for the mini-corpus, the midi-corpus, and finally the main corpus, are split up into a multitude of subsets according to their distinctive features. The resulting subsetted distance tables are then described statistically to conclude whether the distance tables of the two subsets created by one metadata point are significantly different. As a result, it will first and foremost be possible to determine whether parameter settings and feature selections have any influence on the similarity within a subset and the dissimilarity to another corresponding subset. Additionally, it will be possible to ascertain whether particular parameter settings intensify differences between pairs of subsets. A clearly detectable difference between subsets based on gender, for example, would imply that the parameter setting used favours the detectability of a gender signal.

Summing up, it is the aim of this experiment to confirm or refute the following two null hypotheses:

H_{0_1}*: Parameter settings and feature selections do not affect the p-value of a significance test when comparing distance tables created through random subsetting.*

H_{0_2}*: The significance value of differences between subsets based on distinctive features do not vary based on influences of parameter settings and feature selections.*

As mentioned before, it is unfortunately not possible to use non-binary classes for this experiment, which is why the genre category was omitted. Furthermore, texts with a `NA` value in a certain category had to be excluded from the subset of this category in order to ensure the binarity of options. This means that a text whose author's nationality is unclear will not appear in the subset for nationality, but will be included in all other subsets.

5.1 Approach

Descriptive statistics is not necessarily a tool which is widely used in CLS analyses by itself, even though significance tests are employed regularly to validate results yielded with other methods. In contrast to this, descriptive statistics and significance tests are commonly used in corpus linguistics to compare and contrast corpora and subsets and to assess differences in language usage (see, for example Baayen 2008, Gries 2009, Gries 2015, Lijffijt et al. 2016). As I have adopted Jakobson's definition of literary texts as condensations of natural language and will also examine differences in feature frequencies, it seems justifiable to make use of these methods, even if they are not as established in CLS.

The general idea of stylistic signals is that there are groups of texts that share a certain characteristic which causes them to be significantly different from other groups of texts. But in which way can they be significantly different? Texts that are stylistically similar use, according to the basic principle of stylometric analyses, specific features similarly often. The distance value, which is calculated using these feature counts, is smaller for texts which resemble each other in style. In order for a stylometric signal to be identifiable, it has to be divisive enough to separate texts into groups which are through their inward similarity different to the rest of the corpus. Because statistically, comparing a subset with another subset which, when combined, cover the entire dataset, corresponds to a comparison of the subset with the entire corpus. This means that by testing two subsets against each other, each subset is also tested against the whole dataset. If texts written by women, for example, are significantly different from texts written by men, the female subset can also be claimed to behave significantly different in comparison to the whole corpus.

There are many ways to compute the aforementioned inward similarity; I have chosen the most straight-forward option of the arithmetic mean.[1] For each text in a subset, the mean for all its distances is computed, assuming that texts which have a common style would have similar mean values. Then, these sets of means are statistically compared with the other subset created by the same distinctive feature using the t-test.

The t-test is a parametric statistical test which was chosen due to two reasons. First, it is already established in corpus linguistics (Baayen 2008, Lijffijt et al. 2016) and should therefore also be applicable to literary corpora. Second, it is a suitable test for the kind of data that is examined. Textual data is often described as not normally distributed, which

[1] In `stylcoR`, I have also implemented the use of the median and the standard deviation for this purpose.

would mean that parametric tests, which assume normal distribution, are not a valid choice. However, I have already applied the method of z-scoring, which infers normal distribution. When checking the sets of means using a Shapiro test, the values are therefore normally distributed. Applying a t-test on data which is created from z-scores is therefore possible. Moreover, with the exception of the epistolary subset, all used datasets have more than 30 observations, which is often regarded to be a threshold value for assuming normal distribution (Hogg, Tanis, and Zimmerman 2015, 202).

There are two levels of influence which will be investigated in this experiment. The first one, corresponding to H_{0_1}, is concerned with the general influence of parameter settings and feature selections. In order to be able to scrutinise the connection between these variables and the p-value of a statistical test of subsets, the variables have to be tested in a neutral context. For this reason, I have created two randomly assembled corresponding subsets. By using an ANOVA[2] test, another frequently used parametric test (Baayen 2008, Gries 2015, Gries 2009), it is possible to compare the influence of more than two variables. This means that by comparing the significance values of subsets based on different parameter settings and feature selections, it is possible to assess the influence of each individual variable option. When comparing the uni- and bigrams, for example, it will therefore become clear whether different features lead to different significance values.

Moreover, random subsets' results will be used to contextualise results yielded from non-randomly created subsets. To create subsets according to metadata specifications, the texts in each corpus have to be split up correspondingly. This is achieved by the function `createMetaSubsets`, which divides the original metadata table into smaller tables, each containing those texts that fit one of the binary features. The function creates, for example, a metadata table for all texts by English writers and another one for all non-English writers. Methodically, this step corresponds with the creation of the randomly selected subsets, which also includes the creation of smaller metadata tables. All these metadata tables are then used by the function `filterDistSubsets` to filter a distance table for texts belonging to one of the binary categories. After saving the resulting distance tables as csv-files, they are then used for the extraction of the already discussed mean values.

Using the function `testSignificance`, the means of all subset pairs are tested against each other. To make their evaluation easier, the significance values and all corresponding information are saved in an overview table. Building on this first testing, the 2,160 significance tests are again tested to examine the effect of specific parameter and feature

[2] ANOVA = Analysis of Variance

options on each individual subset. This step mirrors the ANOVA test described above and will show whether certain options in- or decrease the significant differences between metadata subsets.[3]

5.2 Results

The first observation that can be drawn from the results is that the vast majority of subset pairs were significantly different. The overall outcome shows that from the 2,160 subset pairs (360 parameter subsets, 6 subset categories (5x metadata, 1x random)), only 15 are not significantly different. Having a closer look at those outliers helps to recognise first patterns: All of them are subsets built on the category of epistolary texts, all of them use the parameter setting 100 MFF, and none of them is built on the entire corpus, but only the mini- and midi-corpus. However, this also means that the differences detected between the randomly assembled subsets were significant which can be seen as a clear indication that this level of inquiry is not very reliable.

When inspecting the random subsets in detail, several observations can be made. Table 5.1 indicates to which degree of certainty variations of significance values can be linked to settings and selections. Generally, the influence of settings and selections seems to decrease with a growing dataset. For all options but the choice between uni- and bigrams, the degree of certainty describing the influence of these options on the significance levels of detected differences drops consistently in the midi- and the main corpus. The significance value for the difference between uni- and bigrams rises, but never exceeds the confidence level of 0.95. For the main corpus, no option seems to cause any significant changes in the significance values; only the variations of the MFF size come, with a p-value of 0.054, close to any form of considerable impact. Over all corpora, the most important differences in the category of MFFs can be detected between the variations 500-100, 1000-100, and 3000-100. All other variations, i.e. those not involving the option of 100 MFFs, are not significant in any corpus. There are no significant differences between any culling values in any corpora. The same applies to the comparison of ternarisation and normalisation.

Exploring the subsets which are based on metadata categories strengthens the observations made above, but also provides more detailed information. Table 5.2 displays the influence of parameter settings and features in general. The values highlighted in blue represent the significant influences on the significance values of differences between subsets. This

[3] The evaluation of significance with the ANOVA test was carried out with a stand-alone script which is not incorporated in the package. It can be found in the appendix.

	random_mini	random_midi	random_main
tranformation_p_value	0.011001651	0.087753817	0.100818664
normalise-none	0.016576104	0.10842192	0.122986498
ternarise-none	0.062269796	0.224416349	0.244435029
ternarise-normalise	1	1	1
MFF_p_value	0.000339099	0.019519941	0.054376325
500-100	0.002216715	0.048625005	0.091611102
1000-100	0.0021674	0.048586416	0.091611102
3000-100	0.002157677	0.04863358	0.119997416
1000-500	0.999999896	1	1
3000-500	0.999999821	1	1
3000-1000	0.999999999	1	1
cosine-burrows	0.047153446	0.142841055	0.154631128
culling_p_value	1	1	1
50c-20c	1	1	1
80c-20c	1	1	1
80c-50c	1	1	1
bigram-unigram	0.532597495	0.089537726	0.080630328

Table 5.1: Detailed Significance Values of Parameter Settings and Feature Selections in Randomly Selected Subsets

might sound complicated at first, but basically tells us that, for example, the significance of differences between epistolary and non-epistolary texts in the main corpus does not vary significantly in the 360 parameter and feature combinations. The same is true for almost all subsets of the main corpus, with the exception of the novel subset, which is significantly influenced by the MFF size. Thus, the details reinforce the perception that the more observations, i.e. individual texts, a data set has, the less it is dependent on parameter and feature options. On the other hand, this also means that for smaller corpora, the impact of variations in settings is comparatively significant: While the main corpus has only one significant value, the midi-corpus has seven and the mini-corpus even ten.

Concerning the settings, general tendencies described in the discussion of the random subsets also apply here. The MFF size appears to be an important impact for many subsets, as it has a p-value below 0.05 for eleven out of fifteen subsets. Transformations are influential for four subsets, different distance measures for two, and n-gram variations for one. Moreover, the culling value does, as described above, not influence any subsets.

	transformation	MFF	measure	culling	ngram
epistolary_main	0.106027949	0.147947578	0.172603495	0.615533122	0.11741172
epistolary_midi	0.023140022	0.001327083	0.215292045	1	0.031302423
epistolary_mini	0.001302167	3.99057E-06	0.262563333	1	0.244888802
gender_main	0.100820687	0.054378178	0.154632911	1	0.080629078
gender_midi	0.092350938	0.021596212	0.147054618	1	0.086264121
gender_mini	0.004084392	0.000834308	0.028747762	1	0.411785083
nationality_main	0.100799152	0.054358446	0.154613925	1	0.080642391
nationality_midi	0.100769942	0.025249438	0.15458817	1	0.080659186
nationality_mini	0.094248777	0.022454335	0.148772846	1	0.084472696
novel_main	0.476120921	5.24703E-05	0.416516595	0.371001498	0.319356239
novel_midi	0.059121003	0.033991353	0.114508237	1	0.053535792
novel_mini	0.062983701	0.031593432	0.118588137	1	0.066624624
threshold_1815_main	0.100794158	0.054385408	0.154609522	1	0.080611887
threshold_1815_midi	0.090225909	0.023264894	0.145116208	1	0.0810154
threshold_1815_mini	0.006962623	0.000243214	0.037445638	1	0.773333356

Table 5.2: Overview of Influences on the Significance Value of Differences Between Subsets

5.3 Discussion

The results presented above show very clearly that varying parameter settings and feature selections influence the p-values of differences between randomly selected subsets significantly. Thus, H_{0_1} can be refuted. In more general terms, this means that these variations affect the similarity of means within a subset and the dissimilarity to another subset. Especially for smaller text collections from up to 200 texts, the choice of parameters and features can thus be consequential. For more extensive collections, i.e. collections comprising more than 500 texts, these choices seem to have less impact, which could be used as an argument for refraining from using any variations in these cases (cf. Underwood 2019).

An observation that caught my eye immediately is that of the seemingly non-influential culling value. For neither the random subsets nor the metadata subsets, any significant impact is noticeable. This seems surprising, as the culling value is supposed to limit the feature sets quite rigorously. It thus makes sense to have another look at the way varying culling values affect distance tables. For this purpose, I compared a randomly chosen set of distance tables created with the same parameter settings, but different culling values. All distance tables belonging to this set (unigram, burrows-delta, 1,000

MFFs, no transformation) have the same distance values for each text. This suggests that they were compiled on the same data basis. When examining the corresponding tables of frequencies, it becomes clear that they in fact do built on identical features and feature counts. I first interpreted these findings as indicators of a logical error in the implementation and consequently went back to inspect the function `performCulling`. It turned out, however, that the implementation is correct, and that I had an incorrect understanding of the force of impact culling has on the feature selection. For the exemplary settings mentioned above, the feature list comprises 141,381 unigrams before the culling. Of these 141,381 features, 4,386 lay above the culling value of 80%, i.e. occur in more than 80% of all texts in the corpus. This means that in this case, the varying culling level does not cause any differences in the feature lists. In the course of all experiments conducted, only the combination of the main corpus, 3,000 unigrams, and a culling value of 80% led to a change of the feature list. Generally, this can be interpreted to show that for smaller English-language corpora of up to around 200 texts, the importance of culling is negligible.

	gender	nationality	threshold_1815	novel	epistolary
500-100	0.0916136	0.0915867	0.091632996	1	0.9999528
1000-100	0.0916136	0.0915867	0.091624003	1	0.9997641
1979-100	0.6599506	0.6599086	0.659950476	2.36E-05	0.9999948
3000-100	0.1200004	0.1199685	0.12000033	1	0.2086444
1000-500	1	1	1	1	0.9982263
1979-500	1	1	1	2.36E-05	0.9999286
3000-500	1	1	1	1	0.2582551
1979-1000	1	1	1	0.0000236	0.9999999
3000-1000	1	1	1	1	0.1473541
3000-1979	1	1	1	3.158E-05	0.7200189

Table 5.3: Significance Values of MFF Sizes in All Subsets of the Corpus

The only parameter combination affected by culling results in a more limited feature list, which has, as mentioned before, only 1,979 items, and causes the significant differences in the corresponding novel subsets, i.e. the only main corpus subset which is influenced by the MFF size. Differences caused by such an extreme limitation of features are not surprising. It is, however, extremely interesting that significant dissimilarities can only be found in the novel subsets (see Table 5.3). One could argue that this proves that the difference between broader texts genres, i.e. the difference between novels and non-novels, is strongly affected by words of medium frequency which occur in at least 80 percent of all texts in the corpus. Nevertheless, when having another look at one of the affected frequency tables, it appears that this implementation of generic difference

is not as nuanced as hoped and simply differentiates texts according to text length: In 112 texts, which represents roughly 20 percent of the corpus, the last 22 unigrams of the feature list do not occur at all, many of them have zero counts for features from around the position 1,600 onwards. All of these 112 texts are shorter prose texts and their quantitative restriction causes this lack of lexical diversity. The distinction between novels and non-novels is thus not a differentiation based on generic conventions, but solely based on length.

This is also partially true for the midi- and the mini-corpus. Of the 182 texts in the midi-corpus, 27 texts, i.e. around 15 percent, do hardly feature any of the unigrams from position 2,000 onwards and have zero counts for many of the items listed before that. For the mini-corpus, the proportion of similarly behaving texts lays by around 14 percent. Again, the affected texts are shorter prose texts, which means that the differences between novels and non-novels is simply based on the distinction between longer and shorter texts. Nevertheless, when comparing Table 5.4 with the values in Table 5.1, it becomes apparent that the novel subsets behaves differently than the randomly selected subsets. As this means that there are in fact differences in the impact a parameter or feature option has on different metadata subsets, H_{0_2} can be refuted as well.

	novel_mini	novel_midi	novel_main
tranformation_p_value	0.062983701	0.059121003	0.476120921
normalise-none	0.080247326	0.075772482	0.506378362
ternarise-none	0.18294709	0.175948416	0.634503647
ternarise-normalise	1	1	1
MFF_p_value	0.031593432	0.033991353	5.24703E-05
500-100	0.999998837	0.999999992	1
1000-100	0.998736415	0.992984195	1
3000-100	0.06370378	0.061982015	1
1000-500	0.998334902	0.992766719	1
3000-500	0.061441102	0.061553225	1
3000-1000	0.091343673	0.116279244	1
cosine-burrows	0.118588137	0.114508237	0.416516595
culling_p_value	1	1	0.371001498
50c-20c	1	1	1
80c-20c	1	1	0.441022522
80c-50c	1	1	0.441022522
bigram-unigram	0.066624624	0.053535792	0.319356239

Table 5.4: Detailed Significance Values of Parameter Settings and Feature Selections in Novel Subsets

Similarly, the nationality subsets show, when compared to the random subsets, other patterns of influence. For these subsets, the only significant differences are produced by the MFF sizes in the mini- and midi-corpus. But what can be inferred from these values? It can, for example, be claimed that variations of language use connected to regional dialects might not affect function words and high-frequency features. This would be an explanation for the fact that the significant differences are only identified when another MFF size is compared to 100 MFFs. Another explanation could be that the 20 percent culling filters out regional variations of certain terms and leaves only their standard English counterparts. Therefore, a group of texts which do not feature these standard terms might be significantly different to a group of texts using them.

	nationality_min	nationality_mid	nationality_main
tranformation_p_value	0.094248777	0.100769942	0.100799152
normalise-none	0.115684589	0.122932504	0.122964874
ternarise-none	0.23450249	0.244362308	0.244405907
ternarise-normalise	1	1	1
MFF_p_value	0.022454335	0.025249438	0.054358446
500-100	0.054193449	0.059246876	0.091586654
1000-100	0.054161433	0.059246095	0.091586652
3000-100	0.054059515	0.059246101	0.119968508
1000-500	1	1	1
3000-500	1	1	1
3000-1000	1	1	1
cosine-burrows	0.148772846	0.15458817	0.154613925
culling_p_value	1	1	1
50c-20c	1	1	1
80c-20c	1	1	1
80c-50c	1	1	1
bigram-unigram	0.084472696	0.080659186	0.080642391

Table 5.5: Detailed Significance Values of Parameter Settings and Feature Selections in Nationality Subsets

In contrast to the nationality subsets, significant differences of the epistolary, gender, and temporal threshold subsets are attributed to a wide range of setting variations (see section 9.3 for the tables of detailed significance values). These distributions are somewhat similar to those of the randomly compiled subsets. There are three possible reasons for this variety of influences. First, texts belonging to one class of the subsets, i.e. epistolarity, texts by women, and texts written before 1815, are, although they are influenced by their own class similarity, similar to other groups of texts based on other categories. In contrast to length and language variety, which are, as explained above, perhaps the main drivers

for similarity and dissimilarity in the novel and nationality subsets, the quintessence of epistolary texts, female authorship, and publication year are not as easily condensed. This can be interpreted as an indication that these subsets are not as well defined as the others and that the distinctive features which were used in their compilation need to be partitioned in smaller, more concrete, features. Second, it might be possible that the variations are in fact all caused by one specific signal. This would mean that epistolary, gender, and threshold subsets are easily dividable into two classes. It is, however, not very plausible that this in fact applies to all of these subsets, as the random subset has similar significant differences. Therefore, it might be the case that third, there is simply no inward similarity to be detected, and the subsets are thus structurally similar to random selection.

Another valuable lesson learnt from this experiment is that more extensive corpora are not as strongly impacted by parameter variations as smaller corpora. In other words, the smaller a corpus is, the more time should be invested in the selection and evaluation of parameters and features. Especially variations of the MFF size seem to have a considerable impact on to which extent differences in the comparison of groups are significant. For both the mini- and the midi-corpus, the disparity between MFF sizes is significant for all subsets. This can be seen as a possible confirmation of the theory mentioned in the theoretical chapters regarding the idea that some MFF sizes are better suited to detect a certain metadata signal than others. However, it is crucial to point out that these results do not suggest that any option is better than another, but only that the differences between the options are substantial.

In sum, both null hypotheses suggested above can be refuted. Nevertheless, this refutation is not clear-cut and only applies to some parameter settings, some feature selections, and some subsets. For the following two experiments, it is helpful to keep in mind that a change in the culling value will yield varying results in only very few cases. Moreover, it will be interesting to see whether the first observations made about tendencies of metadata subsets can be reproduced with other methods.

6 Experiment 2: Classification

The first experiment has shown that some parameter settings and feature selections influence the significance of differences between subsets and that in the context of metadata subsets, some settings and selections have greater impact than others. It has, however, not yet indicated whether these differences favour or hinder the detection of a metadata signal. The second experiment, a classification, aims to answer this question.

In a classification, a possible stylistic signal does not build on a subset's mean values and their inward similarity and outward dissimilarity. The rules according to which texts are classified as belonging to one class or another depend on the classification algorithm used. In this case, a SVM classifier was implemented, which was chosen because it is regularly used in text classification (Manning, Raghavan, and Schütze 2008, Carstensen et al. 2010). This type of classifier uses the training data to define specific areas in a multi-dimensional space which belong to one of the classes by creating a dimension for each feature. The values that each text has for each feature are projected onto these dimensions. After mapping each texts belonging to the training data onto the multi-dimensional space, the algorithm finds the best possible option to split the training data into two separate groups of classes. Building on this spatial separation, texts belonging to the test data are also projected onto the predefined space and are classified according to in which of these separated sections they are placed.

In contrast to the other experiments, I will not use distance matrices, but z-scores of features for this approach. This is primarily due to the fact that classifiers generally rely on feature scores, but also to examine similarities and differences without having to distinguish between distance measures.

In short, the two following null hypotheses will be addressed with the classification experiment.

H_{0_1}*: Classifying texts according to different metadata categories does not lead to varying accuracy values.*

H_{0_2}: When classifying texts according to a metadata category, the parameter settings and feature selections do not affect the accuracy.

Analogously to the previous experiment, only binary classes will be taken into account for the classification processes.

6.1 Approach

As mentioned above, the data used for this experiment does not consist of distance matrices, but z-scores of features. This means that not all 360 parameter settings have to be taken into account, as the z-scores are not affected by distance measures. Consequently, only 216 z-scores are used for the classification. Each one of them is first merged with a filtered metadata table. As a result, each text is described by feature counts on the one side, and a metadata category, on the other.

Using functionalities provided by the classification package `caret` (Kuhn 2019), the texts are split up into training and test sets, each of which mirrors the proportions of the distribution of categories in the respective corpus. If, for instance, there are almost twice as many texts written by men than by women in the mini-corpus, this ratio is reproduced in both the training and the test data. Then, the training data, which comprises 80 percent of all available texts, is used for training the classifier model. Again, this is achieved by using the helpful functions provided by the `caret` package, which offers an easily implementable k-fold cross-validation option. In this development step, a classifier's reliability is raised: The training set is separated in multiple sections, so-called folds, of the same size. Using each of these sections as test data once, the classifier iterates over the training data multiple times. It can be assumed that this refines the classification, as many more test cases are examined. After the model is used to predict the classes of the test data, the predictions and importance values for the used features are saved. Additionally, a table summarising all predictions, their settings, accuracy, precision, and recall is compiled. All this is done by calling the function `classifySVM`.

To ensure reproducibility, a so-called seed parameter is used for the classification. This parameter is an integer variable which is implemented with the default value of 100. As it is called before the data is split into training and test set, it ensures that no matter when, where, and by whom the function is called, the classification will lead to the same results. For the purposes of hypothesis testing, it might be, however, necessary to run the

classification on a diverging data splitting. If this is the case, any other integer value can be passed to the function, which then results in a varying partitioning.

6.2 Results

Of the 1,080 classification runs, 384 had a higher accuracy value than 80 percent and 186 made accurate classification for even more than 90 percent of their test data. When comparing the accuracy levels of parameter settings for one metadata category statistically, it becomes clear that at least some of them vary significantly.

	tranformation_p_value	MFF_p_value	culling_p_value	ngram_p_value
gender	0.880056785	0.122316435	0.995623654	0.041852293
nationality	0.019425074	0.113575776	0.996980209	0.002388364
threshold_1815	0.222903021	0.287099853	0.997635282	4.15729E-17
novel	0.616481967	1.08112E-05	0.985132044	9.29798E-06
epistolary	0.01659421	0.787019635	1	0.019819724

Table 6.1: Significance Values of Parameter Settings and Feature Selections for the Classification of All Metadata Categories

Table 6.1 shows transparently that a classification's accuracy depends notably on the choice of n-gram size. As expected, the culling value does not influence the classification, as it hardly produces any differences in a feature list. The fact that the generic classifications are influenced by a changing MFF size is also expectable because, as expanded on above, the differentiation between novels and non-novels is intrinsically linked to lexical richness, and thus implicitly text length. It is more surprising that both nationality and epistolarity are categories significantly impacted by z-score transformations.

On a more detailed level, epistolarity is also the category with the highest overall accuracy values with a mean accuracy of 97.96 percent, followed by the differentiation of novels and non-novels (80.56%) and works written before and after 1815 (77.54%). The other categories gender and nationality are correctly assigned only slightly better than chance with mean accuracy values of 66.35 percent and 60.23 percent.

What can be deduced from these results? At first, it seems that the classifier works extraordinarily well when differentiating texts according to the existence or non-existence of a epistolary structure. But this first impression is misleading: A closer look at the classifications shows that the classifier actually defines all texts as non-epistolary. As the

proportion of epistolary texts is very small for each corpus (see Figure 3.3), the results are assumed to be very accurate. This means that the classification of epistolary texts is in fact not outstandingly good, but even completely glosses over the categorical differences between the groups of texts. For all other categories, no such imbalance between a seemingly good accuracy value and less than ideal recall can be identified on such a broad scale.

corpus	metadata_col	accuracy	precision	recall	transformation	MFF	culling	ngram
corpus	novel	0.882882883	0.96429	0.69	none	1000	20c	2
corpus	novel	0.882882883	0.96429	0.69	normalise	1000	20c	2
corpus	novel	0.864864865	0.96154	0.64	normalise	3000	20c	2

corpus	metadata_col	accuracy	precision	recall	transformation	MFF	culling	ngram
midi_corpus	novel	0.916666667	1	0.57	ternarise	500	20c	2
midi_corpus	novel	0.888888889	0.8	0.57	ternarise	1000	20c	2
midi_corpus	novel	0.833333333	1	0.14	none	1000	20c	1
midi_corpus	novel	0.833333333	0.66667	0.29	normalise	1000	20c	1
midi_corpus	novel	0.833333333	1	0.14	ternarise	1000	20c	1
midi_corpus	novel	0.833333333	1	0.14	normalise	1000	20c	2
midi_corpus	novel	0.833333333	1	0.14	none	100	20c	2
midi_corpus	novel	0.833333333	0.66667	0.29	normalise	3000	20c	2

corpus	metadata_col	accuracy	precision	recall	transformation	MFF	culling	ngram
mini_corpus	novel	0.863636364	0.75	0.6	ternarise	3000	20c	1
mini_corpus	novel	0.863636364	1	0.4	ternarise	500	20c	1
mini_corpus	novel	0.863636364	1	0.4	ternarise	3000	20c	2

Table 6.2: Details on Classifications of Novels

Nevertheless, the details for all classification runs show that similar problems occur in classifications for several categories, even though the problematics are not as far-reaching. Table 6.2 present the most accurate ten percent of classification runs for novels, with the exception of already excluded culling duplicates. Especially for classifications based on the mini- and the midi-corpus, the recall value is often very low. This value indicates how sensitive a model is in the detection of classes, i.e. to which extent non-novels are defined as novels. For the main corpus, the recall is better, but still not ideal. Generally, it seems that a bigger MFF size leads to more reliable results when classifying basic generic differences. Moreover, with growing corpus size, the accuracy of combinations relying on bigrams increases. Interestingly, both ternarisation and normalisation are more prominently featured in these top ten percent than the option of no transformation. This can be seen as an indication that another normalisation step, which levels out text length

differences, as well as an amplification of segmentations helps in the detection of novels.

In contrast to the differentiation of novels and non-novels, the classification of texts according to the time of their publication has above-average values for both accuracy and recall (see Table 6.3). All of the top classification runs are based on unigrams and vary notably in the size of the feature list. Again, transformation processes seem to be crucial for a model's performance.

corpus	metadata_col	accuracy	precision	recall	transformation	MFF	culling	ngram
corpus	threshold_1815	0.864864865	0.871287129	0.977777778	ternarise	3000	20c	1
corpus	threshold_1815	0.864864865	0.86407767	0.988888889	ternarise	500	20c	1
corpus	threshold_1815	0.855855856	0.855769231	0.988888889	normalise	100	20c	1
corpus	threshold_1815	0.855855856	0.862745098	0.977777778	ternarise	1979	80c	1

corpus	metadata_col	accuracy	precision	recall	transformation	MFF	culling	ngram
midi_corpus	threshold_1815	0.857142857	0.857142857	0.96	none	100	20c	1
midi_corpus	threshold_1815	0.857142857	0.857142857	0.96	normalise	100	20c	1
midi_corpus	threshold_1815	0.857142857	0.884615385	0.92	ternarise	3000	20c	1

corpus	metadata_col	accuracy	precision	recall	transformation	MFF	culling	ngram
mini_corpus	threshold_1815	0.904761905	0.882352941	1	ternarise	1000	20c	1
mini_corpus	threshold_1815	0.904761905	0.933333333	0.933333333	none	500	20c	1
mini_corpus	threshold_1815	0.857142857	0.875	0.933333333	normalise	100	20c	1
mini_corpus	threshold_1815	0.857142857	1	0.8	ternarise	3000	20c	1

Table 6.3: Details on Classifications of Works Before and After 1815

Similar to classifications of epistolary texts, the majority of classifications of authorial gender and nationality do not capture their subjects very well, with either a low recall or precision value. In the majority of classification runs, hardly any or no differences are identified and all test cases are classified as belonging to only one option of the binary category. There are, however, some exceptions: For the category of gender, comparatively good results and balanced out accuracy, precision, and recall values are achieved when combining 100 or 500 MFF sizes with ternarisations or normalisations for both uni- and bigrams and when using the main corpus. For the distinction between English and non-English authors, the unigrams seem to be slightly better suited than bigrams, even though the classification quality for nationalities varies extremely over all corpora.

6.3 Discussion

In general, the results presented above indicate that some metadata categories can be classified with a better accuracy than others. Therefore, H_{0_1} can be refuted, particularly because the mean accuracy values range from only slightly over 60 to almost 98 percent. As indicted in Table 6.1, parameter settings and especially feature selections have a significant impact on classification results, which is why H_{0_2} can be rebutted as well. On a more detailed note, it is extremely important to consider not only the accuracy, but also precision and recall values when evaluating classifications, especially when there is an uneven partitioning of classes.

As the results for epistolary texts are extremely biased and thus misleading, I will not discuss and interpret them here. For similar reasons, I will exclude the category of nationality, since most of the texts are classified as English and hardly any overlap between the very few correctly identified non-English texts can be determined, which can also be seen as a sign for unreliability. I will begin by analysing the few balanced out results for gender and will then go on to describe classifications of the categories of the novel and the temporal threshold.

The best results for gender are produced when combining 100 or 1,000 MFFs with normalisations and the main corpus. The fact that the best results are yielded when using the main corpus might be seen as proof that using more data for the training processes ameliorates the performance. Of the 110 texts in the test set, 86 were correctly classified. Out of the 24 misclassified texts, 22 were texts by female authors classified as texts by male authors, and two were written by men, but recognised as being written by women. This ratio changes slightly for the classifications based on 1,000 MFF, as 84 texts are correctly attributed and 26 are misclassified. Of the in sum 47 texts which are misclassified as texts by male authors in both runs, 20 are constant over both classification runs. Despite this overlap, there seems to be no recognisable pattern for the misclassifications. Many different genres are involved, ranging from sensational to children's fiction. Moreover, there is no clear tendency for a certain time period to be misjudged, and even the texts by male authors which were misinterpreted—Arthur Conan Doyle's "The Adventure of the Priory School" (1905), Thomas Holcroft's *The Adventures of Hugh Trevor* (1794), and George MacDonald's *David Eldginbrod* (1863)—do not have much in common. There are some female authors, like George Eliot, Margaret Oliphant, and Mary Shelly, whose texts are always classified as male, but as the underlying systematics is not clear, it might be too far-fetched to draw any conclusions from this.

For the classification of the publication phase, I expected the misclassifications to be systematic in the sense that texts published around 1815 would tend to be incorrectly attributed. This is not the case for the mini-corpus, as of the eight texts incorrectly classified in the best three runs for the mini-corpus, three were published before 1750 and four after 1850. In the midi-corpus, there is already a noticeable change to this dynamic, because the publications years of 14 out of 15 misclassified texts fall in between 1750 and 1850. A similar observation can be made in the main corpus: Of the 46 incorrectly classified texts, 15 were published before 1750, 30 between 1750 and 1850, and one after 1850. Even if these findings are only built on a sample basis, it seems that texts which were published around 1800 are more likely to be incorrectly attributed. For future uses of similar methods, an automatised evaluation of these results would help to solidify such theories. Moreover, it becomes clear that the mini-corpus might be—despite using methods like k-fold cross-validation—too small to produce reliable results in a supervised learning setting.

Using the same evaluation regime for the classification of novels, I examined the most successful classification runs to find patterns in the attribution. In contrast to the example described above, the mini-corpus is very much in line with the two more extensive text collections. Many of the texts incorrectly classified in the mini-corpus are consistently misclassified in all other corpora. Even more interestingly, the method of classification does not seem to be as length-sensitive as descriptive statistics. There are several comparatively short texts, as for example, Rudyard Kipling's short children's stories, which are classified as novels. The same applies to stories by Arthur Conan Doyle and Joseph Sheridan LeFanu. In some cases, one could suspect that texts are classified according to authorial style, as, for example, Mary Edgeworth's story "To-Morrow" (1804) is classified in line with two of her other works, the novels *Castle Rackrent* (1800) and *Ormond* (1817), and Thomas Hardy's "Interlopers at the Knap" (1888) is also identified as a novel and consequently as belonging to the predominant genre in Hardy's oeuvre.

Overall, it is crucial to think about what classifications are supposed to do: They are designed to classify individuals according to a pattern taken from a sample. The sample thus enormously influences what the model defines as, for example, novels and non-novels. If the training set contains many more novels by one author than by any other, there is an intrinsic bias in the model. In my case, the main corpus contains numerous novels by Walter Scott and a multitude of stories by Rudyard Kipling and Arthur Conan Doyle because they were mentioned very frequently in the secondary sources. This means that it might be necessary to define training sets more rigorously to achieve unbiased results—an approach which would go into the direction of what Underwood describes as sampling (2019, 173-184).

7 Experiment 3: Networks

The final experiment introduces two new categories for the examination: specific genre and authorship. In contrast to both the descriptive statistics approach and the classification, the method of network analysis allows for more nuance and does not define a text as belonging to one group or another. Texts are connected to others on the basis of similarity and can then be described and interpreted according to their position in the resulting network. Thus, a more multi-layered analysis is possible. The network approach is therefore less formalised than both other methods used. As a consequence, the results are, however, also not as definite as the differentiation between significant and not significant differences or correct and incorrect classifications.

Again, a slightly different definition for stylometric signals has to be chosen. In the context of network analysis, stylometric signals are neither based on a subset's mean values, nor on the position a texts assumes in a multi-dimensional space, but on the way a text connects with others. Consequently, this is the most structuralist approach, as the structural context of a corpus does not only inform and influence the results, but is fundamental for their interpretation. When, for instance, one specific parameter setting shows a text written before 1815 in a cluster with other texts from this period, and other parameter settings do not display similar structures, it can be assumed that the texts are hold together by a corresponding stylistic signal. One of the advantages of this approach is that many different metadata categories can be used in the interpretation of one network. Therefore, it will be possible to ascertain whether there is more than one category which might be the cause for a particular grouping behaviour.

With this specific focus, the experiment's aims can be summarised in the following two null hypotheses:

H_{0_1}: *Networks based on filtered distance matrices do not change in structure because of parameter variations and feature selections.*

H_{0_2}: Some parameter variations and feature selections emphasise groupings based on specific metadata categories.

7.1 Approach

Network analysis is, similar to descriptive statistics, a method not necessarily widely used in CLS projects. In the domain of stylometry, Eder, Rybicki & Kestemont (2016) have been the first to implement them; Matt Erlin has already used them some years earlier in the context of a topic modeling project (2014). Even though Eder, Rybicki & Kestemont's method is extremely informative, it is not suitable for my experimental design, as they use information gathered in the computation of consensus trees for their networks. By doing so, they produce networks of reliable connections. In contrast to their study, which focuses on these strong links, I am more interested in weaker links between texts that vary across a range of parameter settings and feature selections. My approach will mirror a method proposed by Thomas Weitin (2019), which uses distance matrices filtered by the so-called Simmelian Backbone algorithm (Nick et al. 2013). Contrary to this contribution, I will not apply a filtering algorithm, but will introduce filtering parameters.

First, all distance tables are, one by one, read in by the function `createLinksNodes`. Then, the values in the distance tables are inverted. This has to be done as distance tables and network links are normally based on opposite logics: In a distance table, a small value equals strong similarity, in a network, small values are usually interpreted as weak similarities. Thus, all distance values are transformed (1\x) to level out this differences. The smallest values in the original distance table are now the biggest. As the distance between a text and itself is always 0—which is why distance tables are adjacency matrices—this computation introduced NA values through the division by 0. In order to be able to use the distance table, all these values have to be transformed back to 0.

Following this transformation, the filtering method is chosen. I have implemented both a nearest neighbours and a percental cut-off method. The nearest neighbours method, implemented in the function `getNearestNeighbours`, finds the *n* nearest neighbours, i.e. texts with the highest similarity values. All other values are replaced by 0. The percental cut-off method determines all values which fulfil the condition of belonging to the highest *n* percent of all values. Again, all other values are replaced by 0. These replacements are crucial for the next step, in which a network is compiled. For the purpose of creating a base network, the `igraph` package is used (Csardi and Nepusz 2006), which has a specific function (`graph.adjacency_matrix`) for creating graphs from distance,

i.e. adjacency, matrices. Each remaining value in the distance matrix is interpreted as a weighted link between the two concerning texts, zeros are interpreted as the absence of links.

The resulting `igraph` object is then converted into a `networkD3` object (Allaire et al. 2017), which offers on the one hand more options for a dynamic visualisation, but also facilitates the splitting of the network object into links and nodes. By saving these two elements separately in csv-files, it is also possible to work with the generated networks in programs like Visone (Brandes and Wagner 2004) and Gephi (Bastian, Heymann, and Jacomy 2009), as lists of links and nodes belong to the standard input options for both applications.

By importing the metadata table and extracting the relevant category, attributes can be added to the nodes created in the previous step. These attributes are then used for the colouring of the nodes, for which a colour palette provided by the `RColorBrewer` package (Neuwirth 2014) was implemented. Depending on the class a given text was ascribed to in the metadata table, the node representing this text is coloured accordingly. As a layout algorithm, I have chosen the option of the `forceNetwork`, which is already implemented in the `networkD3` package and which contracts nodes based on their links' weights. Since the similarity value is defined as the link's weight, this seems to be a suitable layout because it underlines similarities between individual texts visually. The resulting coloured network is then saved as a dynamic html-file.

This process is repeated for each available metadata category. Consequently, for each of the 120 combinations of parameters and features, 24 differently coloured networks are created. The category of authorship is not used for the colouring, as this would results in a quite confusing display of at least 73 different colours. For this reason, the detection of authorship clusters has to be conducted manually.

7.2 Results

In sum, 8,640 networks were created automatically (120 parameter and feature combinations, three corpora, six metadata categories, and two filtering methods with two settings each). This is, of course, too much data to be presented here in detail. I will therefore choose settings which have proven to be insightful in the previous experiments as starting points and will follow general tendencies of the network structures to detect possible sub-clusters. Nevertheless, I will incorporate as many variations and categories

as possible to paint a full picture of the results and to ensure that my interpretations are built on a solid basis.

The two categories which yielded the most reliable classification results in the previous experiment are those of the temporal threshold and the novel. Both of them had their highest accuracy values for the mini-corpus for combinations including the ternarisation of unigrams. I will begin the discussion of the results with Figure 7.1, which presents one of these highly accurate settings for novel classifications (ternarised z-scores of the 3,000 most frequent unigrams). The network displays the texts in three main clusters, two of which are, in comparison, denser. Neither of the clusters can be attributed to one class only. The more compact clusters on the right and at the top can, however, be described as predominantly novel clusters.

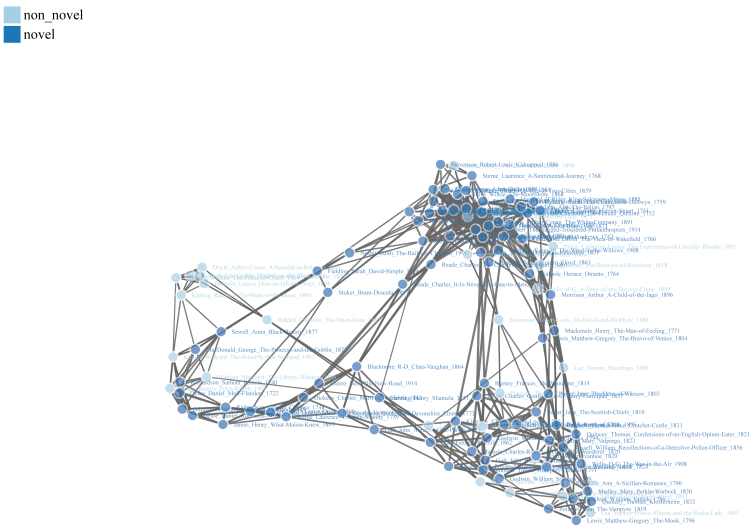

Figure 7.1: Network Based on the 3,000 Most Frequent Unigrams in the Mini-Corpus (Ternarised, Burrows Delta, 6 Nearest Neighbours, Novels)

A comparison with the next network, Figure 7.2, which varies only in the feature selection of bigrams, shows a quite drastic change in the network's outline. The three clusters in the unigram network seem to be pulled apart in its bigram counterpart, so it can be assumed that bigrams create greater and thus more easily detectable differences between groups of

texts. The more intense segmentation has also lead to a more explicit separation of novels, which are predominantly found in the right and the middle cluster, and non-novels, which gather in the small segment on the left.

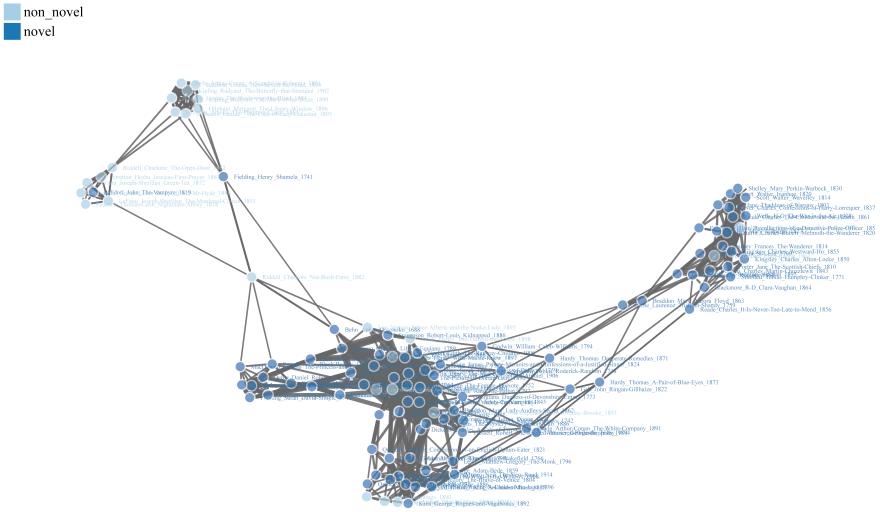

Figure 7.2: Network Based on the 3,000 Most Frequent Bigrams in the Mini-Corpus (Ternarised, Burrows Delta, 6 Nearest Neighbours, Novel)

When comparing the first network with Figure 7.3 and Figure 7.4, it seems like the segmentation is also caused by the ternarisation of z-scores. The two networks show how normalised and unchanged z-scores impact the outline of the network. The normalised network appears to be more homogeneously connected, whereas the network without transformation can be claimed to show already a tendency of separating into three groups. In the normalised network, no individual cluster is constructed; the separation of generic differences does hardly take place. Nevertheless, there is a tendency for non-novel texts to appear in the upper half of the circle. This trend is amplified in the unmodified network, which again shows more distinct clusters. The non-novel texts can mainly be found at the upper parts of the network, where some of them are downright drawn from the centre of the network.

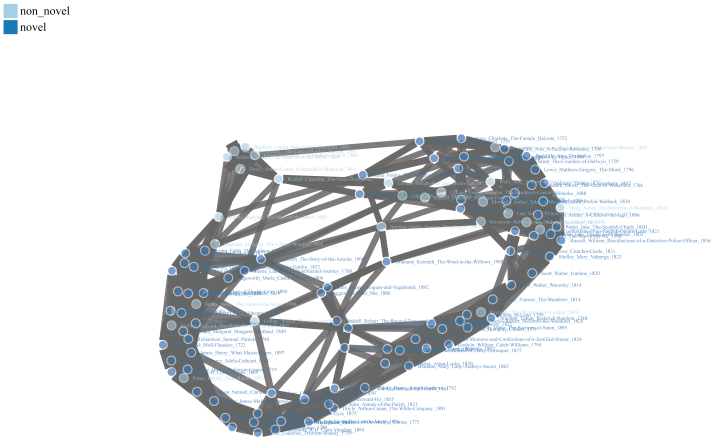

Figure 7.3: Network Based on the 3,000 Most Frequent Unigrams in the Mini-Corpus (Normalised, Burrows Delta, 6 Nearest Neighbours, Novel)

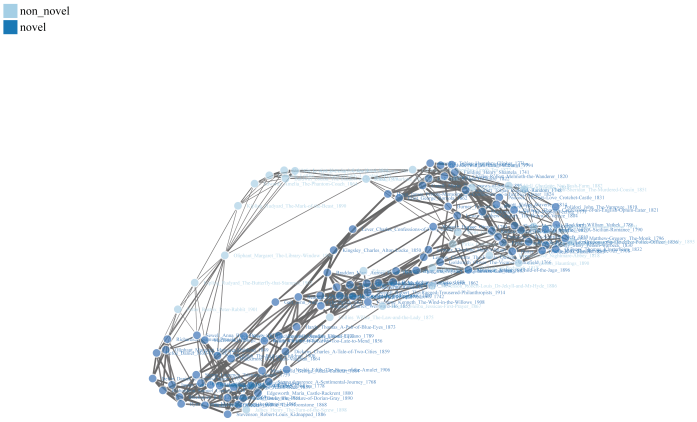

Figure 7.4: Network Based on the 3,000 Most Frequent Unigrams in the Mini-Corpus (No Transformation, Burrows Delta, 6 Nearest Neighbours, Novel)

It is also very intriguing that the more MFFs are used for the creation of a distance table, the more interconnections are created between sub-clusters of a network. The ring-shaped sub-clusters in Figure 7.4 dissolve with the gradual decrease of the MFF size from Figure 7.5 to Figure 7.7. In this gradual change, the main group of non-novel texts first stays at the top position, but is then broken up and scattered across the network.

Figure 7.5: Network Based on the 1,000 Most Frequent Unigrams in the Mini-Corpus (No Transformation, Burrows Delta, 6 Nearest Neighbours, Novel)

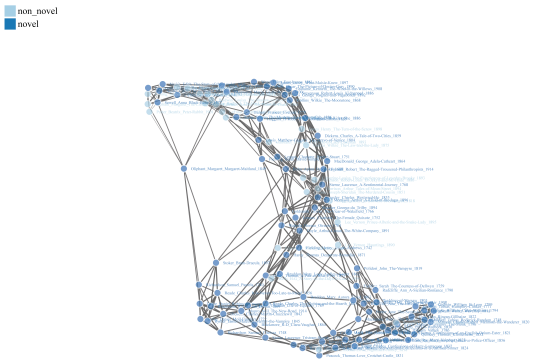

Figure 7.6: Network Based on the 500 Most Frequent Unigrams in the Mini-Corpus (No Transformation, Burrows Delta, 6 Nearest Neighbours, Novel)

Figure 7.7: Network Based on the 100 Most Frequent Unigrams in the Mini-Corpus (No Transformation, Burrows Delta, 6 Nearest Neighbours, Novel)

Exchanging the distance measure has another interesting effect of separation (see Figure 7.8). In comparison to the Burrows Delta, Cosine Delta seems to cause again a more extreme clustering of similar texts, which simultaneously leads to a more intense diverging movement. The non-novels remain relatively scattered. In comparison to all previous networks—with the exception of the bigram network—the formation of opposing segments seems to be more extreme.

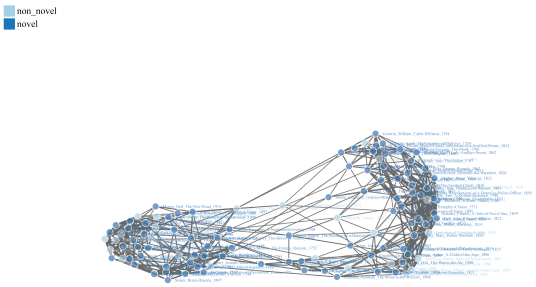

Figure 7.8: Network Based on the 100 Most Frequent Unigrams in the Mini-Corpus (No Transformation, Cosine Delta, 6 Nearest Neighbours, Novel)

7.3 Discussion

The presentation of networks has clearly shown that H_{0_1} can be unequivocally refuted. There are enormous variations in the way networks are formed when even only one parameter is changed. This gives, however, in no way an answer to the question whether variations in parameters and features can be used to created ideal—or at least comparatively good— contexts for the examination of certain metadata categories.

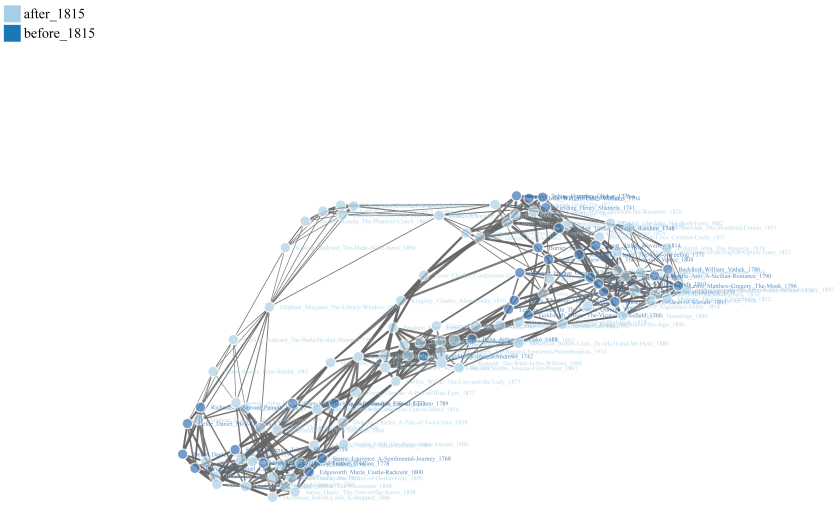

Figure 7.9: Network Based on the 3,000 Most Frequent Unigrams in the Mini-Corpus (No Transformation, Burrows Delta, 6 Nearest Neighbours, Threshold_1815)

Figure 7.4 has already proven to be comparatively suitable for the differentiation of novels and non-novels. The same network is presented in Figure 7.9, expect that here, the category of publication phase is used for the colouring of the nodes. The comparison of these two networks indicates that there is a considerable overlap between the categories, especially regarding the shorter prose texts on the upper part of the network. From a literary history point of view, this observation is not surprising. Since the rise and establishment of the novel as the most dominant literary form begins in the late 17th and early 18th century, it makes sense that the novel clusters are partially also temporal

clusters. Additionally, the identifiable segment of non-novels is also connected by a temporal characteristic, as the majority of them are Victorian sensational and/or mystery stories.

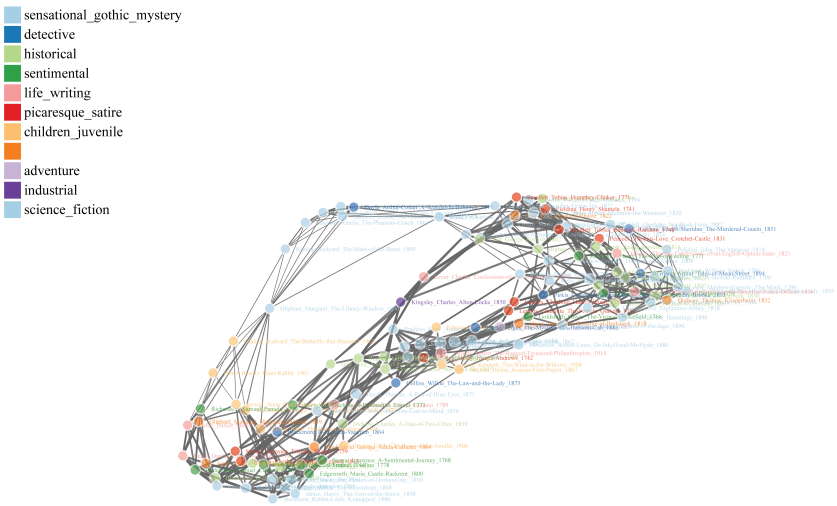

Figure 7.10: Network Based on the 3,000 Most Frequent Unigrams in the Mini-Corpus (No Transformation, Burrows Delta, 6 Nearest Neighbours, Genres)

More specific genres like sensational fiction have therefore a recognisable temporal quality. Thus, it makes sense to map more detailed genre descriptions onto the nodes of this network (see Figure 7.10). Even though the genres are mixed, there seem to be some genre clusters which also correspond with the previous observations. Sensational, gothic, and mystery texts occupy, intertwined with some detective fiction, the outer ridge from the top to the lower right. They are connected to a small children's literature cluster by, amongst others, the texts by Rudyard Kipling, who wrote both sensational mysteries and children's stories. This can be seen as a sign that the authorial signal also plays an important role in this network. Authorship connections can also additionally be found between texts by Daniel Defoe and Samuel Richardson, but not between those by Arthur Conan Doyle and Wilkie Collins. This is especially remarkable, as Defoe and Richardson's texts, *Moll Flanders* (1722) and *Robinson Crusoe* (1719), and *Clarissa* (1748) and *Pamela*

(1740) each belong to the same genre, whereas the texts by Doyle and Collins represent different genres.

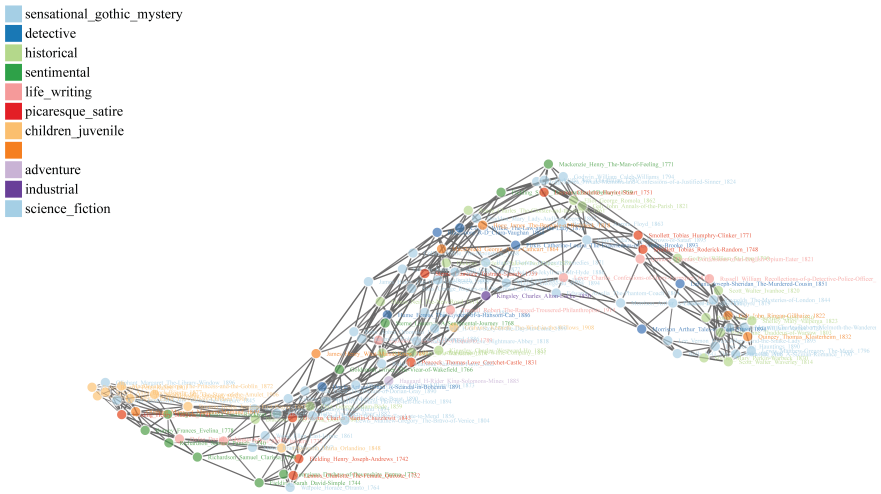

Figure 7.11: Network Based on the 100 Most Frequent Unigrams in the Mini-Corpus (No Transformation, Burrows Delta, 6 Nearest Neighbours, Genres)

Inevitably, these findings raise the questions whether the relationship between genre and authorship changes when other settings are used. For this reason, it makes sense to examine networks based on parameters and features which were previously claimed to produce reliable results in authorship detection. As mentioned in the chapter on methodological foundations, early contributions, but also more recent publications, claim that high frequency features are best suited to detect connections based on authorship (e.g. Mosteller and Wallace 1963, Jockers 2013). Assuming that this can also be applied to networks based on distance tables, the links in Figure 7.11, which is based on 100 MFFs, should connect more authorship couplets than the network in Figure 7.10. This is, however, not the case: In the 3,000 unigram network 16 of 41 possible author connections are established, which is the exactly same number as in the 100 unigram network. Nevertheless, there is some variation of affected texts, as for the 3,000 MFF network, 11 out of 16 of these text couplets also share the same genre, which is only true for eight out

of 16 couplets in the 100 MFFs network.

Because of a lack of detectable influence of this setting, I have tested the theory with another setting which is regularly claimed to achieve better results in authorship detection: Cosine Delta (Evert, Proisl, Vitt, et al. 2015, Jannidis et al. 2015, Büttner et al. 2017). Figure 7.12 shows the Cosine version of the network in Figure 7.10, i.e. 3,000 unigrams with no transformation.

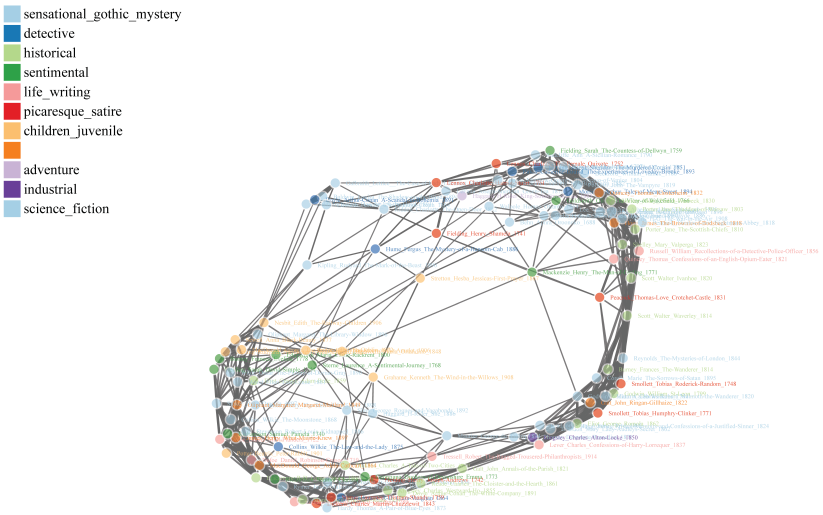

Figure 7.12: Network Based on the 3,000 Most Frequent Unigrams in the Mini-Corpus (No Transformation, Cosine Delta, 6 Nearest Neighbours, Genres)

In this network, 19 of the 41 possible authorship couplets were formed. Evidentially, this cannot be understood as an absolute proof that using Cosine Delta creates more authorship sensitive networks. In spite of that, the increase can be seen as an indication that there are actual differences in the quality of authorship recognition between distance measures. Additionally, 14 of the 19 couplets share both the same author and the same genre. Thus, one could argue that using more features in combination with Cosine Delta might be better suited to capture authorial connections than using a very small feature set. This is due to the fact that the inclusion of only 100 MFFs seems not to be able to show links based on a connection of authorship and genre.

To come back to the network in Figure 7.10, it appears that some genres, by closer inspection, build combined clusters. Picaresque novels (in red) are often linked to sentimental novels (dark green), which they often parody. Together, they build something like the second row to the aforementioned sensational and detective fiction on the edge of the network and help to link them with a small collection of historical fiction. Moreover, there is another accumulation of sentimental novels in the left cluster. This is also the cluster which comprises all three represented epistolary novels. In combination with the two novels by Defoe and *Life of Equiano*, a fictionalised autobiography by Equiano Olaudah, this section can be interpreted as being influenced by the personal and more immediate perspective, both thematically and grammatically, which these texts offer.

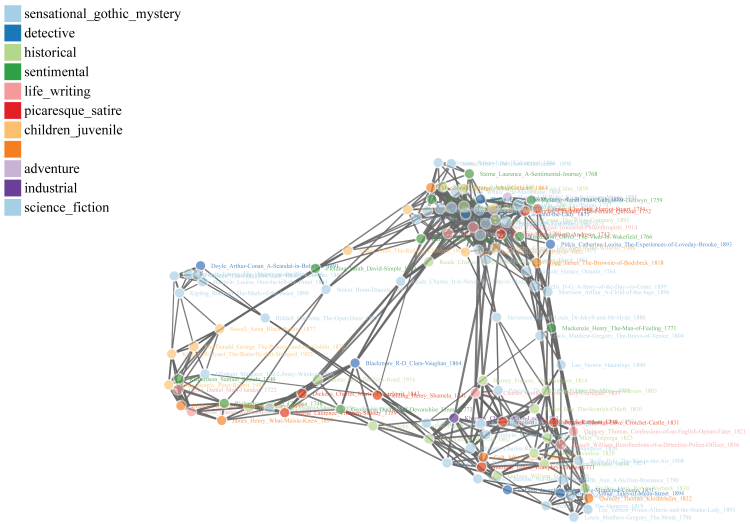

Figure 7.13: Network Based on the 3,000 Most Frequent Unigrams in the Mini-Corpus (Ternarised, Burrows Delta, 6 Nearest Neighbours, Genres)

Building on the insights gained in the presentation of the results, some variations might help to differentiate the network even further. Ternarisations and bigrams have proven to have a divisive influence on networks, which is also apparent in Figure 7.13 and Figure 7.14, which correspond to Figure 7.1 and Figure 7.2, but use genres as attributes for the node colouring. One could assume that stronger contrasts would actually lead to a quantitative rise of distinct genre clusters, but this does not seem to be the case for neither

networks. There are some detectable genre groups, particularly in the bigram network, but the pattern is not more recognisable than in Figure 7.10. I would even argue that the network based on unigrams and no transformations offers a more accurate representation of how genres blend and merge. These subtler connections seem to be lost through the usage of ternarisations and bigrams. Some additional networks (see section 9.4), which are based on the same parameters and filtered with the percental cut-off method, support this claim, as the segmentation and polarisation of clusters accelerates from the base version and the ternarised network to the ternarised bigram network.

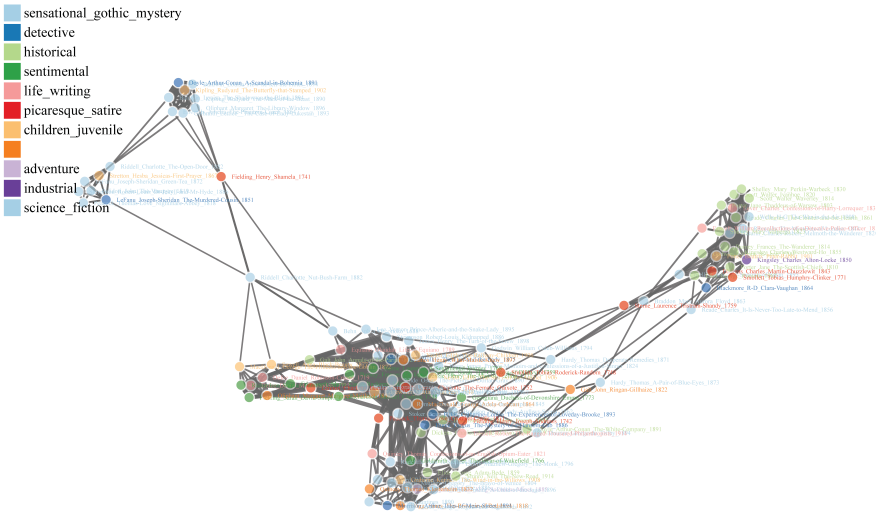

Figure 7.14: Network Based on the 3,000 Most Frequent Bigrams (Ternarised, Burrows Delta, 6 Nearest Neighbours, Genres)

It thus seems that the settings in Figure 7.10 create enough segments to enable a differentiation of genres without losing too much nuance by promoting the formation of more extreme clusters. I have therefore decided to use these settings for my examination of the midi- and the main corpus. The midi-corpus network in Figure 7.15, with its additional 72 texts, is already difficult to interpret because of the added nodes. Being still located at the network's periphery, the sensational short fiction has now moved to the lower part of the network and flows into the collective cluster of life-writing, epistolary, and domestic

texts on the right. Combined with some sentimental novels, this segment could again be seen as being held together by individualised and more direct mediation of plot and content. An additional temporal facet seems to connect the upper right and the upper left, as it features predominantly texts from after 1815.

Figure 7.15: Network Based on the 3,000 Most Frequent Unigrams in the Midi-Corpus (No Transformation, Burrows Delta, 6 Nearest Neighbours, Genres)

The problem of interpretation applies to an even larger extent to the corpus network, which features all 561 texts. There seems to be a division between centre and periphery, which corresponds to the differentiation of longer and shorter texts: Almost all texts which do not belong to the densest central cluster are shorter texts. Moreover, the fringe also contains some dominant author clusters for Kipling and Doyle, whose texts represent a considerable proportion of the non-novel subset. Interestingly, there are also some local hubs for several genres, as, for example, historical novels, sentimental and picaresque novels, and sensational fiction in the central cluster.

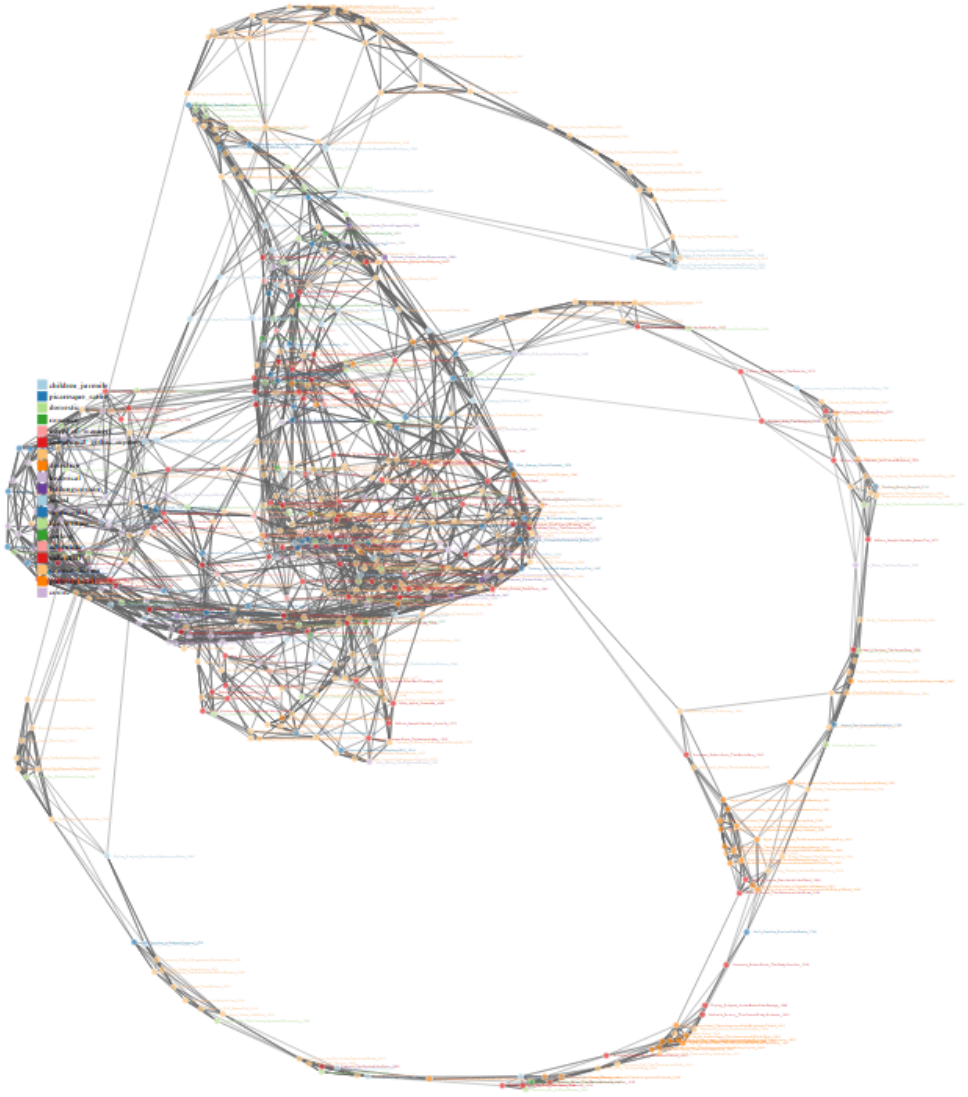

Figure 7.16: Network Based on the 3,000 Most Frequent Unigrams in the Main Corpus (No Transformation, Burrows Delta, 6 Nearest Neighbours, Genres)

By now, it has become clear that detecting and interpreting clusters in large networks manually is extremely challenging and unfortunately, not very reliable and precise. There are community detection algorithms which could be applied here, but which are not yet implemented in `stylcoR`. Depending on manual selection is a major drawback of my approach, especially as it fosters the hyperactive pattern recognition that I have criticised before. In contrast to the more extensive corpora, it seems justifiable to identify clusters manually in the smallest corpus, as it is still small enough to compare and thus validate theories about connections between texts.

Thus, while H_{0_1} can be refuted, H_{0_2} can only be rejected in part. This partially rebuttal is also limited to the mini-corpus, as for the other corpora, more elaborate analyses would have to be involved to be able to test theories of cluster formation. Moreover, it has become very apparent that even though there are some parameter settings and feature selections which are inclined to support the establishments of links based on one specific metadata category, there seems to be not one singular perfect settings for the identification of relationships based on one category.

8 Conclusion

The experiments conducted have proven decidedly that stylometric analyses are highly influenced by parameter settings and feature selections. In all three experimental set-ups, these differences have been detected, even though the intensity of impact varied. What the experiments also exemplified is that quantitative analyses and their results need to be scrutinised and contextualised to be able to draw sound conclusions from them. Generally, some combinations of parameter settings and feature selections seem to clarify and highlight similarity connections based on certain metadata categories. Nevertheless, none of the combinations singled out a specific metadata category as being exceptionally easy to detect.

Beginning with the first experiment, the results have shown that even in randomly selected subsets, significant differences are produced when using diverging parameter settings and feature selections. Several metadata categories notably changed the distribution of these significant differences, limiting them to one option or expanding them to more parameters and features. Generally, the corpus size has been proven to play an important role when assessing influences of parameter settings and feature selections, as the main corpus subsets have hardly been affected by changing options. Importantly, due to the very specific and limited methodical set-up, this experiment can only demonstrate whether the influence of a parameter or feature variation is relevant for the outcome. As a result, it might be best to employ it as a first explorative step in a multi-level approach to decide which kind of parameter and feature options should be used in another analysis.

In contrast to this, the classification has revealed more about how well a specific category can be automatically identified. Besides specific information on how accuracy values and parameters and features are linked, the exploration of classification results has highlighted that some metadata categories achieve good accuracy values only due to over-generalisation. Contrary to these categories, others, as for example the differentiation of publication phase, could not only be classified with more balanced accuracy and recall values, but also showed a systematic scheme in their misclassifications. For other cases, as

for example, the classification according to gender and generic form, no apparent pattern of misclassifications could be identified. A more rigorous evaluation of these results might, however, reveal some sort of logic behind these attributions.

For the purpose of classifications, parameter settings and feature selections which create more contrast or store more information, like ternarisation and bigams, often led to better accuracy values. When using these parameter and feature combinations for networks, it is visualised how they cause structural changes. The usage of both ternarisation and bigrams resulted in more segmented networks, which break up texts into more specific clusters. In contrast to this, networks based on unigrams without transformation tend to be more fine-grained. The question of which parameter and feature combination might be suited best to illustrate a certain metadata category is thus intrinsically linked to how much detail is needed for the description of said category. For a multi-class category like genre, nuance might be more important than the formation of very explicit sub-clusters.

Several systematic flaws were detected in the course of the experiments. There are certainly some experimental designs which yielded seemingly good results but failed to capture the essence of a metadata category. These outcomes were predominantly caused by a lack of sensitivity. In other words, this means that the model did not take into account enough observations or was not supplied with features of sufficient quality to be able to detect patterns of similarity and dissimilarity. One of these cases is the classification of epistolary texts. Other set-ups did produce reliable results, but only by falling back on very simplistic differentiations of groups, which are not connected with stylistic text-intrinsic qualities. The statistical difference between novel and non-novel subsets was, for example, partially defined by text length. Finally, there are some categories which seem to be fundamentally intertwined: A literary genre always has a temporal factor and a text's style can always be influenced by both its author and its genre.

Coming back to the motto of my thesis, the image of signal and noise thus has to be adapted. Some ill-fitted features which influence the outcome of a stylometric analysis can in fact be described as noise, as they produce biased models. Nevertheless, for the disambiguation of stylistic signals, the terminology seems to be not as fitting, as the different stylistic qualities of a text do not obscure each other, but can rather be described as overlapping each other. Some parameter and feature combinations seem to amplify certain signals, but do not simultaneously reduce all others.

Some methodical improvements could be implemented to help to better grasp the systematics of these signals. On a fundamental level, it seems advisable to revise the choice and design of distinctive features, as some have proven to be more valuable than others. Keeping in mind Jakobson's definition of distinctive features, these categories do not need

to cover all possible variations in a corpus but should only implement those which cause an interpretable difference between texts. For the selection of parameters and features, descriptive statistics and significance tests can, as aforementioned, be used to assess possible effects. Because of the high number of individual results, more thorough strategies for summarisation and evaluation have to be developed to ensure that all patterns and tendencies can be detected. This especially applies to the networks, for whose analysis community detection algorithms need to be performed to guarantee objectiveness. From a technical point of view, some of the implemented functions have to be restructured to enhance their efficiency. By improving the package's functionality, issues caused by the introduced limitation of the feature list could be solved, as well.

Lastly, further improvements can also be achieved by better incorporating a point already mentioned in Da's critque of CLS projects. Due to the theoretical and methodological focus of my thesis, literary historical interpretations of the results were only partially supplied. In order to be able to bridge the gap between literary studies and CLS, but also to heighten the understanding of the results, a balance between quantitative methods and their evaluation and literary analysis has to be found.

Bibliography

[1] Mark Algee-Hewitt and Mark McGurl. “Between canon and corpus: six perspectives on 20th-century novels”. In: *Stanford Literary Lab* Pamphlet 8 (2015). URL: https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.

[2] Joseph J. Allaire et al. *networkD3: D3 JavaScript Network Graphs from R*. 2017. URL: https://CRAN.R-project.org/package=networkD3.

[3] Shlomo Argamon, Jonathan Fine, and Rachel Anat Shimoni. “Gender, Genre, and Writing Style in Formal Written Texts”. In: *Text* 23 (Dec. 2003). DOI: 10.1515/text.2003.014.

[4] Shlomo Argamon and Shlomo Levitan. “Measuring the usefulness of function words for authorship attribution”. In: *Proceedings of the 2005 ACH/ALLC Conference*. 2005, pp. 1–3.

[5] Rolf Harald Baayen. *Analyzing linguistic data: a practical introduction to statistics using R*. OCLC: ocn166626226. Cambridge, UK ; New York: Cambridge University Press, 2008.

[6] Roland Barthes. “Science vs Literature”. In: *Twentieth-century literary theory: a reader*. New York, N.Y: Macmillan, 1997, pp. 94–98.

[7] Mathieu Bastian, Sébastien Heymann, and Mathieu Jacomy. “Gephi: An Open Source Software for Exploring and Manipulating Networks”. In: *ICWSM* 8 (2009), pp. 361–362.

[8] Chris Beausang. “‘Some thoughts on Nan Z. Da’s ‘The Computational Case Against Computational Literary Criticism’ or; ‘Against Articles Beginning with the word ‘Against’’”. In: *Medium* (Mar. 2019). URL: https://medium.com/@differengenera/some-thoughts-on-nan-z-17a12445eb3f.

[9] Kenneth Benoit and Adam Obeng. *readtext: Import and Handling for Plain and Formatted Text Files*. 2019. URL: https://CRAN.R-project.org/package=readtext.

[10] Kenneth Benoit, Kohei Watanabe, et al. “quanteda: An R package for the quantitative analysis of textual data”. In: *Journal of Open Source Software* 3.30 (2018), p. 774. DOI: 10.21105/joss.00774. URL: https://quanteda.io.

[11] Douglas Biber. *Variation across speech and writing*. transferred to digit. pr. OCLC: 552099587. Cambridge: Cambridge Univ. Press, 1988.

[12] Douglas Biber and Susan Conrad. *Register, genre, and style*. Cambridge textbooks in linguistics. OCLC: ocn318871987. Cambridge, UK ; New York: Cambridge University Press, 2009.

[13] Kathrine Bode. *Computational Literary Studies: Participant Forum Responses*. Apr. 2019. URL: https://critinq.wordpress.com/2019/04/01/computational-literary-studies-participant-forum-responses-2/.

[14] Ulrik Brandes and Dorothea Wagner. “Visone - Analysis and Visualization of Social Networks”. In: *Graph Drawing Software*. Ed. by Michael Jünger and Petra Mutzel. Springer, 2004.

[15] John Burrows. “‘Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship”. In: *Literary and Linguistic Computing* 17.3 (Sept. 2002), pp. 267–287. DOI: 10.1093/llc/17.3.267. URL: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/17.3.267 (visited on 02/09/2019).

[16] John Burrows. “All the Way Through: Testing for Authorship in Different Frequency Strata”. In: *Literary and Linguistic Computing* 22.1 (Apr. 2007), pp. 27–47. DOI: 10.1093/llc/fqi067. URL: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqi067 (visited on 09/15/2019).

[17] John Burrows. “Questions of Authorship: Attribution and Beyond: A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York”. In: *Computers and the Humanities* 37.1 (2003), pp. 5–32. URL: http://www.jstor.org/stable/30204877.

[18] John Burrows. “Who wrote Shamela? Verifying the Authorship of a Parodic Text”. In: *Digital Scholarship in the Humanities* 20.4 (Nov. 2005), pp. 437–450. DOI: 10.1093/llc/fqi049. URL: http://academic.oup.com/dsh/article/20/4/437/924600/Who-wrote-Shamela-Verifying-the-Authorship-of-a (visited on 09/15/2019).

[19] Judith Butler. *Gender trouble: feminism and the subversion of identity*. Routledge classics. New York: Routledge, 2006.

[20] Andreas Büttner et al. "'Delta' in der stilometrischen Autorschaftsattribution". In: *Zeitschrift für digitale Geisteswissenschaften. text/html Format* (2017). DOI: 10.17175/2017_006. URL: http://zfdg.de/2017_006.

[21] Gerard Carruthers and Liam McIlvanney, eds. *The Cambridge companion to Scottish literature*. Cambridge companions to literature. Cambridge: Cambridge University Press, 2012.

[22] Kai-Uwe Carstensen et al., eds. *Computerlinguistik und Sprachtechnologie*. Heidelberg: Spektrum Akademischer Verlag, 2010. DOI: 10.1007/978-3-8274-2224-8. URL: http://link.springer.com/10.1007/978-3-8274-2224-8 (visited on 10/22/2019).

[23] Robert L. Caserio, ed. *The Cambridge companion to the twentieth-century English novel*. OCLC: ocn291453911. Cambridge, UK ; New York: Cambridge University Press, 2009.

[24] *Chadwyck-Healey Databases*. URL: https://www.proquest.com/products-services/connect/connect_ch.html.

[25] Hugh Craig and Arthur F. Kinney. *Shakespeare, Computers, and the Mystery of Authorship*. OCLC: 876234496. Cambridge: Cambridge University Press, 2012.

[26] Gabor Csardi and Tamas Nepusz. "The igraph software package for complex network research". In: *InterJournal* Complex Systems (2006), p. 1695. URL: http://igraph.org.

[27] Jonathan D. Culler. *Structuralist poetics: structuralism, linguistics and the study of literature*. OCLC: 71287338. London; New York: Routledge ; Taylor & Francis, 2004.

[28] Stuart Curran, ed. *The Cambridge companion to British romanticism*. 2nd ed. Cambridge companions to topics. Cambridge ; New York: Cambridge University Press, 2010.

[29] Nan Z. Da. "The Computational Case against Computational Literary Studies". In: *Critical Inquiry* 45.3 (Mar. 2019), pp. 601–639. DOI: 10.1086/702594. URL: https://www.journals.uchicago.edu/doi/10.1086/702594 (visited on 08/27/2019).

[30] Deirdre David, ed. *The Cambridge companion to the Victorian novel*. 2nd ed. Cambridge companions to literature. Cambridge ; New York: Cambridge University Press, 2012.

[31] Maciej Eder. “Mind your corpus: systematic errors in authorship attribution”. In: *Literary and Linguistic Computing* 28.4 (2013), pp. 603–614. DOI: 10.1093/llc/fqt039. URL: http://dx.doi.org/10.1093/llc/fqt039.

[32] Maciej Eder. “Rolling stylometry”. In: *Digital Scholarship in the Humanities* 31.3 (Sept. 2016), pp. 457–469. DOI: 10.1093/llc/fqv010. URL: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqv010 (visited on 02/09/2019).

[33] Maciej Eder. “Style-Markers in Authorship Attribution: A Cross-Language Study of the Authorial Fingerprint”. In: *Studies in Polish Linguistics* 6 (2011), pp. 99–114. URL: http://search.ebscohost.com/login.aspx?direct=true&db=mzh&AN=2013090078&site=ehost-live.

[34] Maciej Eder. “Visualization in stylometry: Cluster analysis using networks”. In: *Digital Scholarship in the Humanities* 32.1 (Apr. 2017), pp. 50–64. DOI: 10.1093/llc/fqv061. URL: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqv061 (visited on 02/09/2019).

[35] Maciej Eder and Jan Rybicki. “Do birds of a feather really flock together, or how to choose training samples for authorship attribution”. In: *Literary and Linguistic Computing* 28.2 (2013), pp. 229–236. DOI: 10.1093/llc/fqs036. URL: http://dx.doi.org/10.1093/llc/fqs036.

[36] Maciej Eder, Jan Rybicki, and Mike Kestemont. “Stylometry with R: A Package for Computational Text Analysis”. In: *The R Journal* 8.1 (2016), pp. 107–121. URL: https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

[37] Maciej Eder, Jan Rybicki, Mike Kestemont, and Steffen Pielström. “Package ’stylo’”. In: *CRAN* (2019). URL: https://cran.r-project.org/web/packages/stylo/stylo.pdf.

[38] Matt Erlin. “The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731-1864”. In: *Distant readings: Topologies of German culture in the long nineteenth century*. Ed. by Lynne Tatlock and Matt Erlin. 2014, pp. 55–90. URL: http://www.cambridge.org/core/product/identifier/9781571138903/type/BOOK.

[39] Martin Paul Eve. *(Digital) Ways of Looking*. Mar. 2019. URL: https://eve.gd/2019/03/16/digital-ways-of-looking/.

[40] Stefan Evert, Thomas Proisl, Fotis Jannidis, et al. “Understanding and explaining Delta measures for authorship attribution”. In: *Digital Scholarship in the Humanities* 32.suppl_2 (2017), pp. ii4–ii16. DOI: 10.1093/llc/fqx023. URL: http://dx.doi.org/10.1093/llc/fqx023.

[41] Stefan Evert, Thomas Proisl, Thorsten Vitt, et al. “Towards a better understanding of Burrows’s Delta in literary authorship attribution”. In: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Ed. by Anna Feldman et al. Association for Computational Linguistics, 2015, pp. 79–88.

[42] Kate Flint, ed. *The Cambridge history of Victorian literature*. The new Cambridge history of English literature. Cambridge ; New York: Cambridge University Press, 2012.

[43] John Wilson Foster, ed. *The Cambridge companion to the Irish novel*. Cambridge companions to literature. Cambridge, UK ; New York: Cambridge University Press, 2006.

[44] Angela D. Friederici. “The Temporal Organization of Language: Developmental and Neuropsychological Aspects”. In: *Communicating Meaning: The Evolution and Development of Language*. Ed. by B. Velichkovsky and Duane M. Rumbaugh. Hillsdale, Nj: Lawrence Erlbaum Associates, 1996, pp. 173–186.

[45] Stephen Greenblatt and M. H. Abrams, eds. *The Norton anthology of English literature*. 8th ed. OCLC: ocm61229825. New York: W.W. Norton, 2006.

[46] Stefan Th. Gries. “The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models”. In: *Corpora* 10.1 (Apr. 2015), pp. 95–125. DOI: 10.3366/cor.2015.0068. URL: https://www.euppublishing.com/doi/10.3366/cor.2015.0068 (visited on 10/09/2019).

[47] Stefan Th. Gries. “Useful statistics for corpus”. In: 2009.

[48] David Herman, ed. *The Cambridge companion to narrative*. Cambridge companions to literature. OCLC: ocm76794778. Cambridge ; New York: Cambridge University Press, 2007.

[49] Berenike J. Herrmann et al. “Response by the Special Interest Group on Digital Literary Stylistics to Nan Z. Da’s Study”. In: *Journal of Cultural Analytics* (Mar. 2019). URL: https://culturalanalytics.org/2019/05/response-by-the-special-interest-group-on-digital-literary-stylistics-to-nan-z-das-study/.

[50] Robert V. Hogg, Elliot A. Tanis, and Dale L. Zimmerman. *Probability and statistical inference*. Ninth edition. Boston: Pearson, 2015.

[51] Janet Holmes and Miriam Meyerhoff, eds. *The handbook of language and gender*. Blackwell handbooks in linguistics 13. Malden, MA: Blackwell, 2003.

[52] David L. Hoover. "Statistical Stylistics and Authorship Attribution: an Empirical Investigation". In: *Literary and Linguistic Computing* 16.4 (Nov. 2001), pp. 421–444. DOI: 10.1093/llc/16.4.421. URL: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/16.4.421 (visited on 08/27/2019).

[53] Catherine Ingrassia, ed. *The Cambridge companion to women's writing in Britain, 1660-1789*. Cambridge companions to literature. Cambridge, United Kingdom: Cambridge University Press, 2015.

[54] Roman Jakobson. "Linguistics and Poetics". In: *Style in Language*. MIT Press paperback series. Cambridge, Mass: MIT Press, 1960, pp. 350–377.

[55] Roman Jakobson. "Poetry of Grammar and Grammar of Poetry". In: *Verbal Art, Verbal Sign, Verbal Time*. Minneapolis: University of Minnesota Press, 1985, pp. 37–46.

[56] Roman Jakobson. "Two Poems by Puskin". In: *Verbal Art, Verbal Sign, Verbal Time*. Minneapolis: University of Minnesota Press, 1985, pp. 47–58.

[57] Roman Jakobson, C. Gunnar M. Fant, and Morris Halle. *Preliminaries to speech analysis: the distinctive features and their correlates*. 10. print. OCLC: 257542198. Cambridge, Mass: MIT Press, 1963.

[58] Fotis Jannidis et al. "Improving Burrows' Delta – An empirical evaluation of text distance measures". In: 2015.

[59] Matthew Lee Jockers. *Macroanalysis: digital methods and literary history*. Topics in the digital humanities. Urbana: University of Illinois Press, 2013.

[60] Mike Kestemont. "Function Words in Authorship Attribution. From Black Magic to Theory?" In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Gothenburg, Sweden: Association for Computational Linguistics, 2014, pp. 59–66. DOI: 10.3115/v1/W14-0908. URL: http://aclweb.org/anthology/W14-0908 (visited on 08/27/2019).

[61] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. "Automatically Categorizing Written Texts by Author Gender". In: *Digital Scholarship in the Humanities* 17.4 (Nov. 2002), pp. 401–412. DOI: 10.1093/llc/17.4.401. URL: https://doi.org/10.1093/llc/17.4.401 (visited on 05/17/2019).

[62] Max Kuhn. "caret: Classification and Regression Training". In: *CRAN* (2019). URL: https://CRAN.R-project.org/package=caret.

[63] Robin Lakoff. “Language and Woman’s Place”. In: *Language in Society* 2.1 (1973), pp. 45–80. URL: http://www.jstor.org/stable/4166707.

[64] Robin Lakoff and Mary Bucholtz. *Language and woman’s place: text and commentaries*. Rev. and expanded ed. Studies in language and gender. New York: Oxford University Press, 2004.

[65] Jefrey Lijffijt et al. “Significance testing of word frequencies in corpora”. In: *Digital Scholarship in the Humanities* 31.2 (June 2016), pp. 374–397. DOI: 10.1093/llc/fqu064. URL: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqu064 (visited on 10/09/2019).

[66] Devoney Looser, ed. *The Cambridge companion to Women’s writing in the Romantic period*. Cambridge companions to literature. Cambridge, United Kingdom: Cambridge University Press, 2015.

[67] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. OCLC: ocn190786122. New York: Cambridge University Press, 2008.

[68] Gail Marshall, ed. *The Cambridge companion to the fin de siècle*. Cambridge companions to topics. OCLC: ocm76939877. Cambridge ; New York: Cambridge University Press, 2007.

[69] Richard Maxwell and Katie Trumpener, eds. *The Cambridge companion to fiction in the Romantic period*. The Cambridge companion to. OCLC: ocn123029597. Cambridge, UK ; New York: Cambridge University Press, 2008.

[70] Franco Moretti. *Graphs, maps, trees: abstract models for a literary history*. London ; New York: Verso, 2005.

[71] Frederick Mosteller and David L. Wallace. “Inference in an Authorship Problem”. In: *Journal of the American Statistical Association* 58.302 (June 1963), p. 275. DOI: 10.2307/2283270. URL: https://www.jstor.org/stable/2283270?origin=crossref (visited on 02/10/2019).

[72] Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*. 2014. URL: https://CRAN.R-project.org/package=RColorBrewer.

[73] Bobo Nick et al. “Simmelian Backbones : Amplifying Hidden Homophily in Facebook Networks”. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. The Association for Computing Machinery, 2013. URL: https://kops.uni-konstanz.de/bitstream/123456789/25994/1/Nick_259941.pdf.

[74] *Novels Online*. 2016. URL: https://chawtonhouse.org/category/novels-online/.

[75] Michael P. Oakes. "Computer stylometry of C. S. Lewis's The Dark Tower and related texts". In: *Digital Scholarship in the Humanities* 33.3 (Sept. 2018), pp. 637–650. DOI: 10.1093/llc/fqx043. URL: https://academic.oup.com/dsh/article/33/3/637/4652887 (visited on 09/12/2019).

[76] James W. Pennebaker. *The secret life of pronouns: what our words say about us*. Paperback ed. OCLC: 931498035. New York: Bloomsbury, 2013.

[77] Linda H. Peterson, ed. *The Cambridge companion to Victorian women's writing*. Cambridge companions to literature. Cambridge, United Kingdom ; New York: Cambridge University Press, 2015.

[78] Andrew Piper. *Enumerations: data and literary study*. Chicago ; London: The University of Chicago Press, 2018.

[79] Paul Poplawski, ed. *English literature in context*. OCLC: ocn173238916. Cambridge ; New York: Cambridge University Press, 2008.

[80] Jan Rybicki. "Vive la différence: Tracing the (authorial) gender signal by multivariate analysis of word frequencies". In: *Digital Scholarship in the Humanities* 31.4 (Dec. 2016), pp. 746–761. DOI: 10.1093/llc/fqv023. URL: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqv023 (visited on 08/27/2019).

[81] Jan Rybicki and Maciej Eder. "Deeper Delta across genres and languages: do we really need the most frequent words?" In: *Literary and Linguistic Computing* 26.3 (Sept. 2011), pp. 315–321. DOI: 10.1093/llc/fqr031. URL: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqr031 (visited on 09/15/2019).

[82] Ferdinand de Saussure, Wade Baskin, et al. *Course in general linguistics*. OCLC: 695390190. New York: Columbia University Press, 2011.

[83] Ferdinand de Saussure and Peter Wunderli. *Cours de linguistique générale: zweisprachige Ausgabe französich-deutsch mit Einleitung, Anmerkungen und Kommentar*. OCLC: 930951062. Tübingen: Narr, 2013.

[84] Stefan Schöberlein. "Poe or Not Poe? A Stylometric Analysis of Edgar Allan Poe's Disputed Writings". In: *Digital Scholarship in the Humanities* (July 2016), fqw019. DOI: 10.1093/llc/fqw019. URL: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqw019 (visited on 09/12/2019).

[85] Christof Schöch. *Author or genre? Assessing the quality of cluster analysis graphs in two-dimensional classification problems*. Oct. 2012. URL: http://dragonfly.hypotheses.org/148.

[86] Christof Schöch. "Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik". In: *Literaturwissenschaft im digitalen Medienwandel*. Ed. by Christof Schöch and Lars Schneider. 2014, pp. 130–157.

[87] Joanne Shattock, ed. *The Cambridge companion to English literature, 1830-1914*. Cambridge companions to literature. OCLC: ocn436311147. Cambridge, UK ; New York: Cambridge University Press, 2010.

[88] Nate Silver. *The signal and the noise: why so many predictions fail–but some don't*. New York: Penguin Press, 2012.

[89] Ted Underwood. *Distant horizons: digital evidence and literary change*. Chicago: The University of Chicago Press, 2019.

[90] Ted Underwood, David Bamman, and Sabrina Lee. "The Transformation of Gender in English-Language Fiction". In: *Journal of Cultural Analytics* (2018). DOI: 10.22148/16.019. URL: http://culturalanalytics.org/2018/02/the-transformation-of-gender-in-english-language-fiction (visited on 09/12/2019).

[91] Sean G. Weidman and James O'Sullivan. "The limits of distinctive words: Re-evaluating literature's gender marker debate". In: *Digital Scholarship in the Humanities* 33.2 (June 2018), pp. 374–390. DOI: 10.1093/llc/fqx017. URL: https://academic.oup.com/dsh/article/33/2/374/3111279 (visited on 09/12/2019).

[92] Thomas Weitin. "Distanzmaß und Netzwerkanalysen bei einem mittelgroßen literarischen Korpus: Der Deutsche Novellenschatz von Paul Heyse und Hermann Kurz (1871-1876)". In: *forthcoming* (2019).

9 Appendix

9.1 Main Script

```
################################################################################
################################ Instructions ##################################
################################################################################

#### How your data should look like:
####
#### -> The txt files of your corpus should all be in one directory. The
####     variable <path_to_corpus> leads to this directory.
####
#### -> Your metadata table should feature texts as rows and metadata categories
####     as column. One column has to feature the filename of each text (without
####     file extension) so that the metadata can be linked to the files in the
####     corpus.
####

# Variables that have to be altered ############################################

corpus_paths <- c("C:\\Users\\litlab-hiwi\\Documents\\MA-Thesis\\mini_corpus",
                  "C:\\Users\\litlab-hiwi\\Documents\\MA-Thesis\\midi_corpus",
                  "C:\\Users\\litlab-hiwi\\Documents\\MA-Thesis\\corpus")

results_paths <- c("C:\\Users\\litlab-hiwi\\Documents\\MA-Thesis\\results\\mini_corpus",
                   "C:\\Users\\litlab-hiwi\\Documents\\MA-Thesis\\results\\midi_corpus",
                   "C:\\Users\\litlab-hiwi\\Documents\\MA-Thesis\\results\\corpus")


overall_results <- "C:\\Users\\litlab-hiwi\\Documents\\MA-Thesis\\results"

metadata <- read.table("C:\\Users\\litlab-hiwi\\Documents\\MA-Thesis\\distinctive_features_corpus.csv",
                       header = TRUE,
                       check.names = FALSE, sep = ";",
                       stringsAsFactors = FALSE)
# View(metadata)
# Check whether your metadata was read in correctly, you might need to change
# the separator

# What is the column name of your filename column?
file_name <- "file_name"

metadata_cols <- colnames(metadata)
# Drop column containing the file names and non-binary categories!
# In my case, the non-binary category is featured in column 3 (column 1
# features the filename)
metadata_cols_binary <- metadata_cols[-c(1,3)]
# Drop only the filename
metadata_cols_all <- metadata_cols[2:length(metadata_cols)]

################################################################################
# Distance Tables
```

```
##############################################################################

# Parameter Settings ##########################################################

distance_measures <- list("cosine-delta")

MFW <- list(100, 500, 1000, 3000)

culling <- list(20, 50, 80)

n_gram_sizes <- list(2)

zscores_tranformation <- list("none", "normalise")


param_combination <- expand.grid(culling,
                                 MFW,
                                 zscores_tranformation,
                                 distance_measures)


param_combination <- param_combination[!(param_combination$Var3=="ternarise" &
                                         param_combination$Var4=="cosine-delta"),]

# Frequency Tables and Distance Tables ########################################

for (y in 1:length(corpus_paths)) {
  path_to_corpus <- as.character(corpus_paths[y])
  for (x in 1:length(n_gram_sizes)) {
    freq <- createFreqTable(path_to_corpus, n_gram_size = n_gram_sizes[x])
    for (j in 1:nrow(param_combination)) {

      createDistanceTable(path_to_corpus,
                          freq_dist = freq,
                          n_gram_size = n_gram_sizes[x],
                          culling_level = as.numeric(param_combination[j,1]),
                          cut_off = as.numeric(param_combination[j,2]),
                          zscores_transformation =
                            as.character(param_combination[j,3]),
                          distance_measure =
                            as.character(param_combination[j,4]))
      gc()

    }

  }
  rm(freq)
  gc()
}

##############################################################################
# Significance
##############################################################################

# Subsetting ##################################################################

# Metadata Subsets
for (j in 1:length(results_paths)) {
  for (i in 1:length(metadata_cols_binary)) {
    createMetaSubsets(metadata, subset_column = metadata_cols_binary[i],
                      results_paths[j])
  }
}


# Random Subsets
for (j in 1:length(results_paths)) {
  subsets_dir <- paste0(results_paths[j], "\\subsets")
  sub1 <- paste0(subsets_dir, "\\subset_metadata_random_1.csv")
  sub2 <- paste0(subsets_dir, "\\subset_metadata_random_2.csv")

  num <- nrow(metadata)
```

```
  set.seed(100)
  sample_ids <- sample(seq(1, num), num/2, replace = FALSE)
  meta1 <- metadata[sample_ids, ]
  meta2 <- metadata[-sample_ids, ]

  write.table(meta1, sub1, sep = ";", quote = FALSE, row.names = TRUE,
              col.names = NA)
  write.table(meta2, sub2, sep = ";", quote = FALSE, row.names = TRUE,
              col.names = NA)
}

dists <- list.files(results_paths, pattern = "distance_table",
                    recursive = TRUE, full.names = TRUE)

for(n in 1:length(results_paths)) {
  for (m in 1:length(dists)) {
    filterDistSubsets(path_distance_matrix = dists[m],
                      path_results = results_paths[n],
                      file_name_col = file_name)
  }
}



# Significance Test ############################################################

all_dists <- list.files(overall_results, pattern = "dist_table_.+.csv",
                        recursive = TRUE, full.names = TRUE)

results_significance <- data.frame(group1 = character(), group2 = character(),
                        corpus = character(), settings = character(),
                        significant = character(), p_value = numeric())

for (j in seq(1, length(all_dists), by = 2)) {
  t <- testSignificance(all_dists[j], all_dists[j+1], "mean")
  results_significance <- rbind(results_significance, t)
}

results_significance$transformation <- results_significance$setting
results_significance$transformation <- gsub(".+_", "",
                                  results_significance$transformation)

results_significance$MFF <- results_significance$setting
results_significance$MFF <- as.numeric(gsub(".+_(\\d+)MFF_.+", "\\1",
                        results_significance$MFF))

results_significance$measure <- results_significance$setting
results_significance$measure <- gsub(".+_(.+delta)_.+", "\\1",
                           results_significance$measure)

results_significance$culling <- results_significance$setting
results_significance$culling <- gsub(".+_(.+c)_.+", "\\1",
                           results_significance$culling)

results_significance$ngram <- results_significance$setting
results_significance$ngram <- as.numeric(gsub("(.)gram_.+", "\\1",
                           results_significance$ngram))

write.table(results_significance,
            paste0(overall_results, "\\results_significance_tests.csv"),
            col.names = NA)

################################################################################
# Classify
################################################################################

zscores <- list()
for (i in 1:length(results_paths)) {
  results_burrows <- paste0(results_paths[i], "\\burrows-delta")
  z <- list.files(results_burrows, pattern = "zscores",
                  recursive = TRUE, full.names = TRUE)
  zscores <- append(zscores, z)
```

```
}

results_classification <- data.frame(setting = character(),
                                     metadata_col = character(),
                                     accuracy = numeric(),
                                     precision = numeric(),
                                     recall = numeric())

for (j in 1:length(metadata_cols_binary)) {
  for (i in 1:length(zscores)) {
    classification <- classifySVM(path_zscores = zscores[i], metadata,
                                  col_file_name = file_name,
                                  metadata_col =
                                    as.character(metadata_cols_binary[j]))
    results_classification <- rbind(results_classification, classification)
    message(metadata_cols_binary[j], "□no□", i, "□done!")
  }
}

results_classification$transformation <- results_classification$setting
results_classification$transformation <- gsub(".+_", "",
                                              results_classification$transformation)

results_classification$MFF <- results_classification$setting
results_classification$MFF <- as.numeric(gsub(".+_(\\d+)MFF_.+", "\\1",
                                              results_classification$MFF))

results_classification$culling <- results_classification$setting
results_classification$culling <- gsub(".+_(.+c)_.+", "\\1",
                                       results_classification$culling)

results_classification$ngram <- results_classification$setting
results_classification$ngram <- as.numeric(gsub("(.)gram_.+", "\\1",
                                                results_classification$ngram))

write.table(results_classification,
            paste0(overall_results, "\\results_classification.csv"),
            col.names = NA)

#############################################################################
# Network
#############################################################################


all_dists <- list.files(results_paths, pattern = "distance_table.csv",
                        recursive = TRUE, full.names = TRUE)

neigbours <- list(3,6)

param_combination_neigbours <- expand.grid(metadata_cols_all, neigbours)

for (i in 1:length(all_dists)) {
  for (j in 1:nrow(param_combination_neigbours)) {
    graph <- createLinksNodes(path_distance_matrix = all_dists[i],
                              nearest_neighbours = TRUE,
                              cut_off = FALSE,
                              num_neighbours =
                                as.numeric(param_combination_neigbours[j,2]))

    net <- createNetwork(graph,
                         metadata, col_file_name = file_name,
                         metadata_col =
                           as.character(param_combination_neigbours[j,1]))
  }
}

cut_off <- list(1,5)

param_combination_cutoff <- expand.grid(metadata_cols_all, cut_off)

for (i in 1:length(all_dists)) {
  for (j in 1:nrow(param_combination_cutoff)) {
```

```
    graph <- createLinksNodes(path_distance_matrix = all_dists[i],
                              nearest_neighbours = FALSE,
                              cut_off = TRUE,
                              percentage =
                                as.numeric(param_combination_cutoff[j,2]))

    net <- createNetwork(graph,
                         metadata, col_file_name = file_name,
                         metadata_col =
                           as.character(param_combination_cutoff[j,1]))
  }
}

##############################################################################
##############################################################################
```

9.2 ANOVA Evaluation

```
##############################################################################
##############################################################################

significant <- read.csv("results_significance_tests.csv", sep = "□", row.names = 1)

subsets <- list("gender_1", "nationality_1", "threshold_1815_1", "novel_1", "epistolary_1", "random_1")

significant_in_corpus <- subset(significant, corpus == "corpus")

for (i in 1:length(subsets)) {

  sig <- subset(significant_in_corpus, subset1 == as.character(subsets[i]))

  sig$transformation <- as.factor(sig$transformation)

  aov_result_transformation <- aov(sig$significance.p.value ~ sig$transformation)
  sum_transformation <- summary(aov_result_transformation)[[1]][["Pr(>F)"]][1]
  names(sum_transformation) <- "tranformation_p_value"
  x <- TukeyHSD(aov_result_transformation)
  x_transform_difference <- x$`sig$transformation`[,1]
  x_transform <- x$`sig$transformation`[,4]

  transform <- c(sum_transformation, x_transform_difference, x_transform)

  sig$MFF <- as.factor(sig$MFF)

  aov_result_MFF <- aov(sig$significance.p.value ~ sig$MFF)
  sum_MFF <- summary(aov_result_MFF)[[1]][["Pr(>F)"]][1]
  names(sum_MFF) <- "MFF_p_value"
  x <- TukeyHSD(aov_result_MFF)
  x_MFF_difference <- x$`sig$MFF`[,1]
  x_MFF <- x$`sig$MFF`[,4]

  MFF <- c(sum_MFF, x_MFF_difference, x_MFF)

  sig$measure <- as.factor(sig$measure)

  aov_result_measure <- aov(sig$significance.p.value ~ sig$measure)
  sum_measure <- summary(aov_result_measure)[[1]][["Pr(>F)"]][1]
  names(sum_measure) <- "measure_p_value"
  x <- TukeyHSD(aov_result_measure)
  x_measure_difference <- x$`sig$measure`[,1]
  x_measure <- x$`sig$measure`[,4]

  measure <- c(sum_measure, x_measure_difference, x_measure)

  sig$culling <- as.factor(sig$culling)

  aov_result_culling <- aov(sig$significance.p.value ~ sig$culling)
```

```
  sum_culling <- summary(aov_result_culling)[[1]][["Pr(>F)"]][1]
  names(sum_culling) <- "culling_p_value"
  x <- TukeyHSD(aov_result_culling)
  x_culling_difference <- x$`sig$culling`[,1]
  x_culling <- x$`sig$culling`[,4]

  culling <- c(sum_culling, x_culling_difference, x_culling)

  sig$ngram <- as.factor(sig$ngram)

  aov_result_ngram <- aov(sig$significance.p.value ~ sig$ngram)
  sum_ngram <- summary(aov_result_ngram)[[1]][["Pr(>F)"]][1]
  names(sum_ngram) <- c("ngram_p_value")
  x <- TukeyHSD(aov_result_ngram)
  x_ngram_difference <- x$`sig$ngram`[,1]
  x_ngram <- x$`sig$ngram`[,4]

  ngram <- c(sum_ngram, x_ngram_difference, x_ngram)


  overall <- c(transform, MFF, measure, culling, ngram)
  ls[[i]] <- overall
}


overall <- as.data.frame(do.call(rbind, ls))
rownames(overall) <- subsets

write.table(overall, "mini-corpus_sig.csv", sep = "␣", col.names = NA)

################################################################################
################################################################################
```

9.3 Additional Detailed Significance Values of Parameter Settings and Feature Selections in Subsets

	epistolary_mini	epistolary_midi	epistolary_main
tranformation_p_value	0.001302167	0.023140022	0.106027949
normalise-none	0.004657645	0.040374289	0.130359941
ternarise-none	0.00748604	0.07465761	0.247945855
ternarise-normalise	0.900083971	0.979202773	0.999879333
MFF_p_value	3.99057E-06	0.001327083	0.147947578
500-100	0.000175127	0.009766868	0.999952796
1000-100	6.1854E-05	0.00442472	0.999764113
3000-100	3.98301E-05	0.005767697	0.208644383
1000-500	0.99359282	0.994413344	0.998226278
3000-500	0.982439998	0.998280373	0.258255107
3000-1000	0.999542934	0.999798963	0.147354073
cosine-burrows	0.262563333	0.215292045	0.172603495
culling_p_value	1	1	0.615533122
50c-20c	1	1	1
80c-20c	1	1	0.66966281
80c-50c	1	1	0.66966281
bigram-unigram	0.244888802	0.031302423	0.11741172

Table 9.1: Detailed Significance Values of Parameter Settings and Feature Selections in Epistolary Subsets

	gender_mini	gender_midi	gender_main
tranformation_p_value	0.004084392	0.092350938	0.100820687
normalise-none	0.006793914	0.113567098	0.122988739
ternarise-none	0.033686624	0.231583919	0.244438048
ternarise-normalise	1	1	1
MFF_p_value	0.000834308	0.021596212	0.054378178
500-100	0.006480323	0.052705168	0.091613636
1000-100	0.003999744	0.05245886	0.091613636
3000-100	0.003268797	0.052459795	0.120000412
1000-500	0.998791736	0.999999998	1
3000-500	0.996656794	0.999999998	1
3000-1000	0.999917429	1	1
cosine-burrows	0.028747762	0.147054618	0.154632911
culling_p_value	1	1	1
50c-20c	1	1	1
80c-20c	1	1	1
80c-50c	1	1	1
bigram-unigram	0.411785083	0.086264121	0.080629078

Table 9.2: Detailed Significance Values of Parameter Settings and Feature Selections in Gender Subsets

	threshold_1815_mini	threshold_1815_midi	threshold_1815_main
tranformation_p_value	0.006962623	0.090225909	0.100794158
normalise-none	0.010977669	0.111191595	0.12295934
ternarise-none	0.046901081	0.228288266	0.244398453
ternarise-normalise	1	1	1
MFF_p_value	0.000243214	0.023264894	0.054385408
500-100	0.002107337	0.058192904	0.091632996
1000-100	0.001524433	0.054418234	0.091624003
3000-100	0.00151906	0.054413765	0.12000033
1000-500	0.999703198	0.999992945	1
3000-500	0.999693503	0.999992919	1
3000-1000	1	1	1
cosine-burrows	0.037445638	0.145116208	0.154609522
culling_p_value	1	1	1
50c-20c	1	1	1
80c-20c	1	1	1
80c-50c	1	1	1
bigram-unigram	0.773333356	0.0810154	0.080611887

Table 9.3: Detailed Significance Values of Parameter Settings and Feature Selections in Threshold_1815 Subsets

9.4 Additional Networks Produced With a Percental Cut-Off

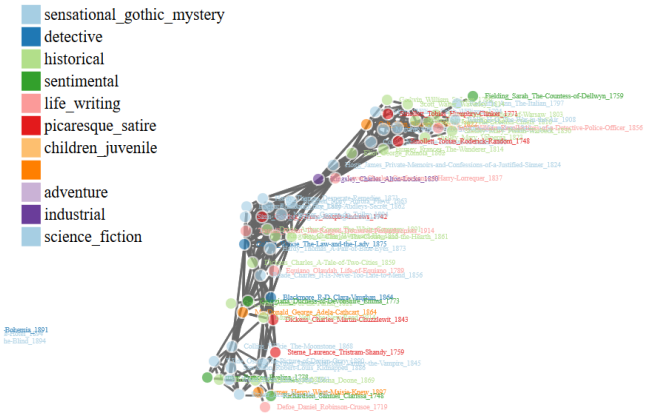

Figure 9.1: Network Based on the 3,000 Most Frequent Unigrams in the Mini-Corpus (No Transformation, Burrows Delta, 5% Cut-Off, Genres)

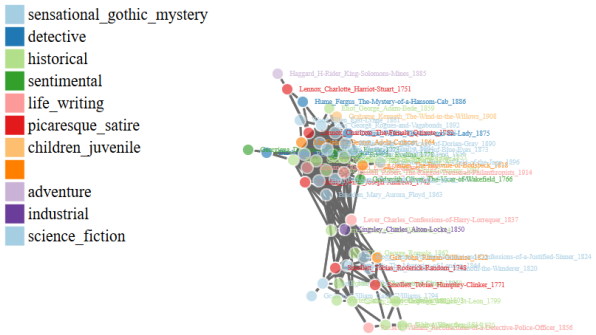

Figure 9.2: Network Based on the 3,000 Most Frequent Unigrams in the Mini-Corpus (Ternarised, Burrows Delta, 5% Cut-Off, Genres)

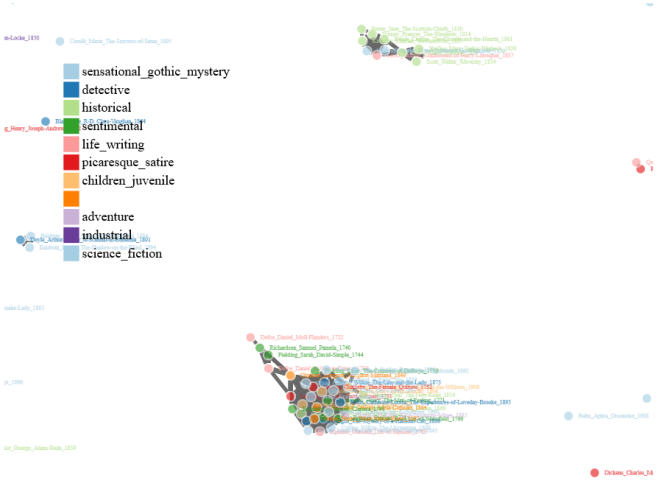

Figure 9.3: Network Based on the 3,000 Most Frequent Bigrams in the Mini-Corpus (Ternarised, Burrows Delta, 5% Cut-Off, Genres)